METHODOLOGY FOR ANALYZING AT-SEA DIVE BEHAVIOUR OF A MARINE MAMMAL

by

Ruth Joy

M.Sc., Simon Fraser University, 2002

B.Sc., University of Victoria, 1996

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

Doctor of Philosophy

in the

Department of Statistics and Actuarial Science
Faculty of Science

SIMON FRASER UNIVERSITY
Summer 2013

APPROVAL

Name: Ruth Joy

Degree: Doctor of Philosophy

Title of Thesis: Methodology for analyzing at-sea dive behaviour of a marine mammal

Examining Committee:

Dr. Tim Swartz, Professor
Chair

Dr. Richard Routledge, Professor
Senior Supervisor

Dr. Dave Campbell, Assistant Professor
Supervisory Committee

Dr. Derek Bingham, Associate Professor
Internal External Examiner

Dr. Loveday Conquest, Professor
School of Aquatic & Fishery Sciences,
University of Washington
External Examiner

Date Approved:

August 29, 2013

Abstract

The population of northern fur seals (*Callorhinus ursinus*) in the Pribilof Islands, Alaska has declined dramatically during the past 35 years. Arresting the decline of the species requires an understanding of their foraging behaviour at sea and is particularly important for those adult females whose foraging success is also linked to pup survival. We propose an augmented state space methodology for studying behavioural patterns using high-resolution movement time series. We show how non-stationary time series models that describe systems for whom parameters evolve slowly over time relative to the state dynamics can be estimated at relevant time scales for behavioural inference. This framework allows us to relate the time-varying parameter estimates of an auto-regressive system model to the seal's at-sea behavior. The at-sea behaviour states of eleven lactating female northern fur seals were then matched, spatially and temporally, to a set of environmental variables, some of which were averages that represented the oceanic conditions over a large spatial area. The mismatch of scale between seal behaviour and the spatial variables was accounted for by applying an error-in-covariate Bayesian hierarchical model. Using this approach, we were able to link together northern fur seals that went to disparate regions of the eastern Bering Sea, with widely variable information about their underlying environmental fields into a single model. This application of a hierarchical model relates changes in identifiable behavioural states of the northern fur seal to changes in the Alaska commercial groundfish industry over a diurnal foraging cycle. The methodology described in this thesis is adaptable for analyzing any type of high-resolution movement data on marine predators, and will allow for the characterization of other at-sea behaviours as well as other descriptors of pelagic habitat and foraging success.

To Cecil, and all those who spent time with him.

"The journey is the reward."

— Old Chinese Proverb.

Acknowledgments

I am grateful to Pamela Lestenkof and Andrew Trites for sharing their animal tag data with me. Pamela's fieldwork was conducted under the authority of Marine Mammal Protection Act permit Nos. 782-1708 and 1045-1713. Financial support for this project was provided by the National Marine Mammal Laboratory and by the North Pacific Marine Science Foundation to the North Pacific Universities Marine Mammal Research Consortium. I acknowledge the Natural Sciences and Engineering Research Council of Canada and the Hakai Network for Coastal People, Ecosystems and Management for funding my studies at Simon Fraser University.

I am deeply indebted to my senior supervisor Rick Routledge for his guidance and support, but most of all for keeping his sense of humour over countless drafts of this thesis. As well, I am indebted to Brian Battaile and Tiphaine Jeanniard du Dot for our many discussions about northern fur seal behaviour, and to Mike Dowd for his friendship and help in understanding the finer points of state-space models.

Many thanks to the faculty and staff of the Department of Statistics and Actuarial Science of Simon Fraser University for providing me a wonderful environment for graduate studies. Thank you also to my fellow graduate students for being good company and for weathering the growing pains of graduate work with me. I would also like to thank the people-of-the-bunker in the Strand Hall Annex for the positive working atmosphere, and for their encouragement during the writing of this thesis.

Finally, I thank my family for their constant support and loving impatience that I be "done-already".

Contents

List of Tables

List of Figures

Chapter 1

Introduction

1.1 Thesis Introduction and Motivation

The eastern Bering Sea is a model system in which to explore relationships between the biological and physical environment and the foraging pattern of top marine predators, such as the northern fur seal (*Callorhinus ursinus*). This region of ocean is a highly productive ecosystem that is characterized by both on-shelf and off-shelf dynamics, such as those that drive prey productivity. The northern fur seal is a wide-ranging marine predator that feeds opportunistically on pelagic species found along long, at-sea foraging paths. The recent and ongoing development of increasingly sophisticated tags capable of recording extensive information when placed on such individuals is generating valuable, but demanding research opportunities.

This study is largely motivated by the conservation needs of the species. Recently, concern about the status of northern fur seal populations has increased, particularly in the Pribilof Islands, Alaska, the site of the largest breeding population for the species (Rice 1998). The Pribilof Island population of northern fur seals has declined dramatically during the past 35 years, and continues to decline without any obvious reason identified (Towell et al. 2006). In particular, pup production has decreased about 50% since the 1970s, with a continuing drop of about 6 or 7% per year. Arresting the decline of the species requires an understanding of the at-sea behaviour and foraging strategies of the northern fur seal (Antonelis et al. 1997), and is particularly important for those adult females whose foraging success is also linked to pup survival. The northern fur seal was listed as "*depleted*" in June 1988 under the US Marine Mammal Protection Act, and as "*vulnerable*" (to extinction) in

2008 by the International Union for the Conservation of Nature (Gelatt and Lowry 2008).

The methods and implementations in this thesis are built around a data set collected from lactating northern fur seals to which tags were attached during the summer (pupping) season. Considerable effort has been invested around the world in attaching tracking devices to many marine species in the hopes of being able to identify at-sea behaviour and, in particular, important influences on feeding behaviour. These tracking devices capture three-dimensional movement at very high-resolution and can then, at least in theory, be linked to the marine ecosystem through which they swim as it varies over both space and time. Full exploitation of the information on at-sea movement is not straightforward, and is an emerging and very active area of research in marine ecology. To date, it is widely recognized that advances in tag engineering have outpaced the developments in statistical methodology (Schick et al. 2008).

The tags used in this study to capture fine-scale movement were Driessen and Kern dead reckoner tags that record measurements every two seconds for the length of the tag deployment. These tags were applied with glue to the backs of a set of lactating northern fur seals at Reef Rookery on the south shore of St Paul Island, in the Pribilof Islands of Alaska (Figures 1.1, 1.2). Using the data retrieved from the tags, we show how to extract meaningful information about animal at-sea behaviour from fine-scale movement data. This thesis addresses the shortfall in statistical methodology for tagged animal data, and provides a framework for incorporating three-dimensional movement data from a collection of animals into models that provide inference about how their behaviour adapts to dynamic changes in their biological and physical environment.

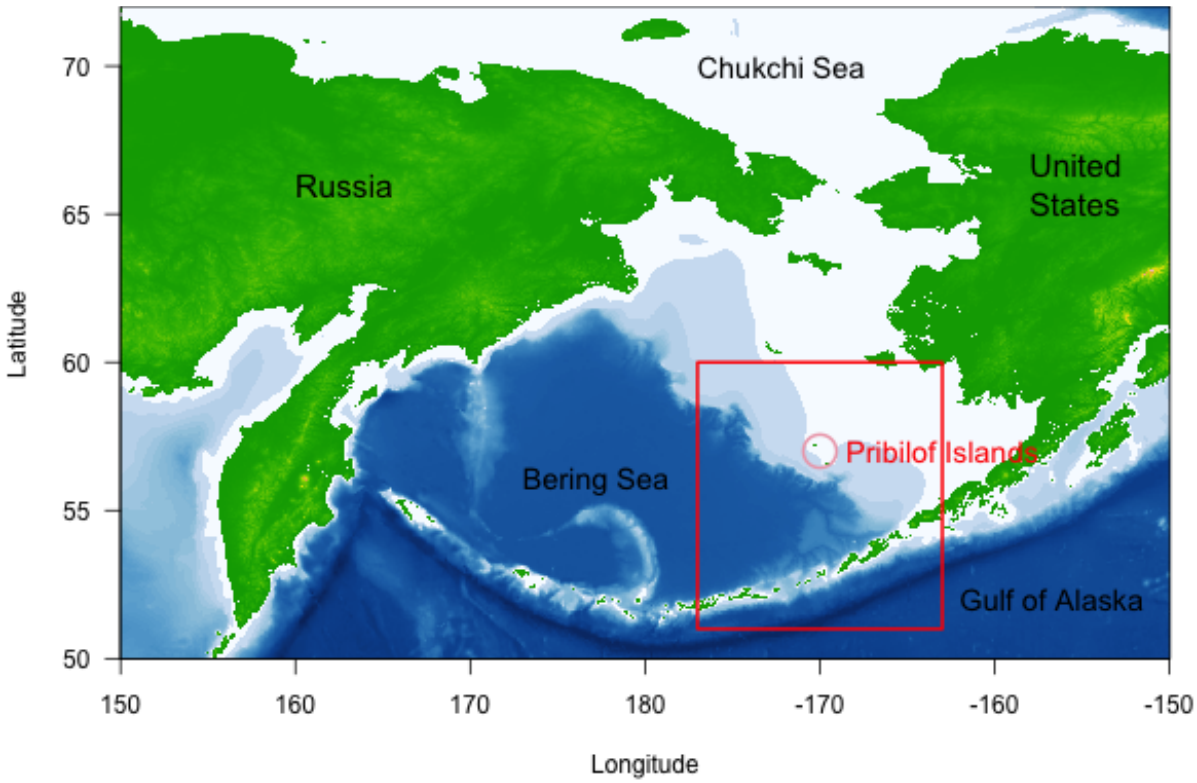

Figure 1.1: Location of the Pribilof Islands in the Bering Sea, Alaska, which are home to the female northern fur seals that are the focal population in this thesis.

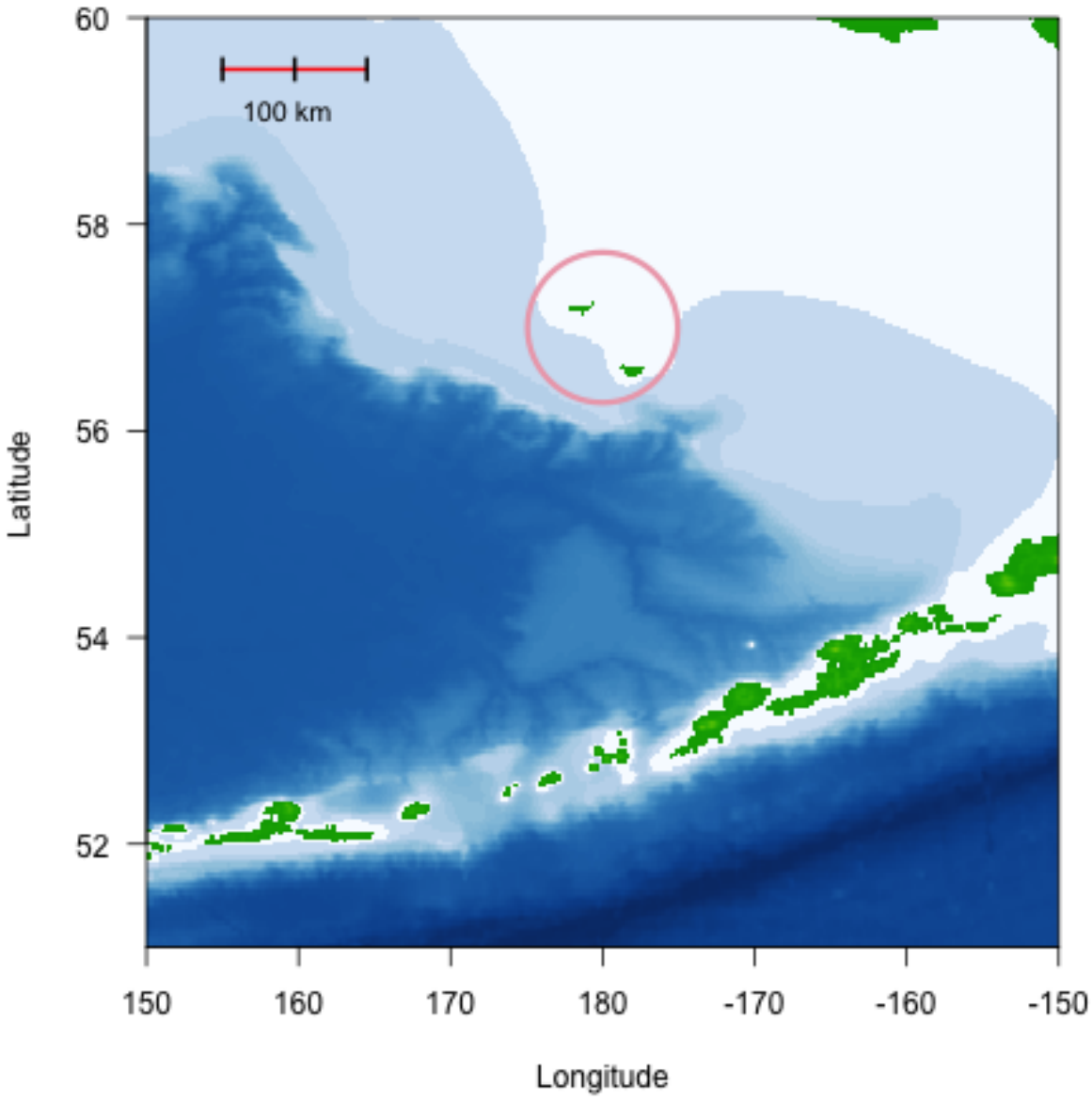

Figure 1.2: St Paul Island (NW in this figure) and St George Island (SE) that constitute the major islands of the Pribilof Islands of Alaska. Reef Rookery is located on a peninsula that extends off the southern coast of St Paul Island. The intensity of the blue colour represents bathymetry of the surrounding ocean. Adjacent waters are considered "on-shelf", and are typically less than 200 m deep, "off-shelf" water starts at 1500 m, and extends to depths greater than 7000 m in the Aleutian Trench located to the bottom right of this figure.

1.2 Overview of Thesis

1.2.1 Non-linear parameter estimation in sequential data

We begin Chapter 2 by introducing the state space model, and show how to represent a physical system (or state) that dynamically evolves over time. We show how the state space model provides a general statistical framework for analysing data from a time series by providing sequential estimates of the state, while accounting for errors in the system and observation processes. The state space approach, implemented via particle filters and sequential Monte Carlo methods, is convenient for handling multivariate data as well as nonlinear-non-Gaussian processes (Ristic 2004).

In our application where the parameters as well as the state need to be estimated simultaneously, the state vector can be reconfigured by appending the parameter vector to the state vector. By augmenting the state vector with the parameter vector, and by adding perturbation to those estimates, the augmented state space model can then be solved using the same sequential Monte Carlo methods as for regular state space models. The final development in this chapter is to show how non-stationary time series models that describe systems for whom parameters evolve slowly over time relative to the state dynamics can be estimated. We reformulate the parameter estimation problem to focus on estimation at time scales that can be directly linked to more meaningful time frames for behavioural inference. These iterated filtering methods allow for uncertainties in the data and the process model, but also include estimates and uncertainties in the parameters.

1.2.2 An application of the particle filter for estimating animal movement parameters

Chapter 3 describes an implementation of the iterated filtering method applied to a 24 hour section of a northern fur seal's at-sea foraging track. We define the system as a discrete time-dependent problem, and frame the process model as a 2^{nd}-order ordinary differential equation model. The differential equation process model allows for time series that show damping, oscillations and random walk characteristics, all of which are interpretable characteristics of animal movement.

This example implementation highlights issues which are expected to arise in any analysis of high-resolution movement data. They include

- the identification of an appropriate movement model,
- the specification of well-defined values for the magnitude and form of the process and observation errors;
- the estimation of the minimal number of identifiable parameters, and determining parameters off-line wherever possible; and
- choosing the appropriate time scale for parameter estimation to best match emergent behaviour.

1.2.3 Hierarchical Bayes methods for modelling uncertainty in at-sea animal behaviour

Increasingly, the framework for understanding an ecological or environmental problem is being confronted with issues surrounding the integration of multiple data sources, multiple data types, and the modelling of hierarchical relationships.

In Chapter 4, we introduce the hierarchical Bayesian model as a means to account for multiple sources of uncertainty in the measurement of and the understanding of the fur seal's at-sea environment. We specifically develop a multinomial hierarchical modelling framework in which coherent behaviours can be identified and extracted from the continuous-movement, state space solutions and linked through a linear process model to imprecise measures of environmental covariates. The imperfect measures of covariate data are explicitly incorporated using a Berkson error model (Berkson 1950), to avoid biased estimates of our regression estimates.

The flexibility afforded by Bayesian methods, and the flexibility of modern computational methods within the Bayesian paradigm mean that this is an extendable and adaptable framework suitable for other environmental data, other processes and other species.

1.2.4 An application of a hierarchical Bayesian model of seal behaviour linked to biological and physical fields

In Chapter 5, we use the hierarchical Bayesian approach of Chapter 4 to link together northern fur seals that went to disparate regions of the eastern Bering Sea into a single model that informed us about the expected behaviour of the population of maternal, female northern fur seals at Reef Rookery, on St. Paul Island. Our analysis focussed on three

typical fur seal behaviours, and how these behaviours are associated with time of day, and a set of environmental data. The approach and analysis presented here is to use hierarchical Bayesian modelling to bring together coarse estimates of the environment, and the dynamics of fur seal behaviour through identification of variable-length segments of coherent behaviours. In this, we have successfully shown that northern fur seals actively forage more at night, and in areas where their preferred prey species, juvenile walleye pollock (*Theragra chalcogramma*), appears to be more abundant.

1.3 Literature Cited

Antonelis, G.A., E.H. Sinclair, R.R. Ream, and B.W. Robson. 1997. Inter-island variability in the diet of female northern fur seals (*Callorhinus ursinus*) in the Bering Sea. Journal of Zoology, 242: 435-451.

Berkson, J. 1950. Are there two regressions? Journal of the American Statistical Association, 45(250): 164-180.

Gelatt, T. and L. Lowry. 2008. *Callorhinus ursinus*. *In*: IUCN 2008. IUCN Red List of Threatened Species. Listed as Vulnerable (VU A2b).

Rice, D.W. 1998. Marine Mammals of the World - Systematics and Distribution (Special publications of the Society of Marine Mammalogy special publication 4). Allen Press, Inc., Lawrence, Kansas, 231pp.

Ristic B., S., Arulampalam, and N. Gordon. 2004. Beyond the Kalman Filter: Particle Filters for Tracking Applications. Artech House: Boston, 318pp.

Schick, R.S., S.R. Loarie , F. Colchero, B.D. Best, A. Boustany, D.A. Conde, P.N. Halpin, L.N. Joppa, C.M. McClellan, and J.S. Clark. 2008. Understanding movement data and movement processes: current and emerging directions. Ecology Letters, 11: 1338-1350.

Towell, R.G., R.R. Ream, and A.E. York. 2006. Decline in northern fur seal (*Callorhinus ursinus*) pup production on the Pribilof Islands. Marine Mammal Science, 22(2): 486-491.

Chapter 2

Non-linear Parameter Estimation in Sequential Data

2.1 Introduction

This chapter introduces the state space model to represent a physical system that dynamically evolves over time. The state space model provides a general statistical framework for analysing data from a time series by providing sequential estimates of components from a dynamic system. The state space model uses stochastic differential equations to describe the state's time evolution from time $t-1$ to time t, and an observation equation that accounts for the imperfect observation of the state. The splitting of the time evolution into two modelling equations allows for errors originating both with the state and measurement processes to be understood separately. Through the state model, we can incorporate complex models to describe the evolution of the model state over time, while allowing errors to accumulate therein. The observation model allows for the incorporation of a different kind of error that accumulates through the measurement process. If both functions that govern the state and observation equations of the system are known and linear, with only Gaussian error terms, then the Kalman filter (Kalman 1960) is the maximum likelihood estimate of the state, and also the estimate with minimum mean square error.

Unfortunately, nonlinearity and nonnormality are common in real data sets, and the use of the Kalman filter for these analyses is no longer optimal. In this thesis, we begin by introducing particle filtering methods as an alternative when the assumptions of the Kalman

filter can't be met. Particle filter methods are well established and a popular method for solving state space models. The particle filter presented in this chapter uses sequential Monte Carlo methods that rely on the idea of using samples to characterize probability distributions.

For many analyses, the state space models are not based on well-known equations in which parameter values can be realistically assumed known. Often when using particle filters, the likelihood surface is rough as the realization of that likelihood is dependent on the discrete nature of the particles themselves (Doucet and Tadić 2003). This chapter proposes an alternate approach to maximum likelihood estimation of state parameters that uses a state-augmented approach to the state equation following the methods of Kitagawa (1998).

One key aspect to time series analyses is identifying the appropriate time scale for meaningful inference. Often data are not collected at a time scale that can be directly linked to meaningful inference about a dynamic system. We introduce a non-standard application of particle filter methods using augmented state vectors that allow model parameters to slowly evolve with time. This method uses overlapping windows of times series data to sequentially carry out an iterative filtering algorithm that results in parameters that evolve slowly relative to the state dynamics.

2.2 The State Space Model

A state space model is a way of representing a dynamic system in the time domain as a set of input, output, and state variables. In this thesis, a dynamic system is defined as a rule which describes the time evolution of a model, given a set of parameters relevant to the system. This model has a state consisting of a collection of points in a state space, where those points form a sample from the range of possible values for the state variable. In practice, states of interest may be the kinematic characteristics of a target in a tracking problem (Ristic et al. 2004), the unobserved true signal in signal processing (Liu and Chen 1998), the actual spoken words in speech recognition (Rabiner 1989), the underlying volatility in an economic time series (Pitt and Shepard 1999), the composition and thickness of a semiconductor alloy (Maars 2000), among many others.

In any dynamic system, at least two models are required for inference using the state space framework. One model describes the evolution of the state of the system, while

another relates the observations to the unobserved state. Such a representation is known as the state space form of the system (Harvey 1989). If a time series model is expressed in state space form, several powerful estimation techniques may be applied, such as the Kalman filter. The state variable of interest and equations governing the system may be expressed in a compact form using vector or matrix equations. The general form of a state space model is as follows:

$$\boldsymbol{x}_t = f(\boldsymbol{x}_{t-1}, \theta, \boldsymbol{\eta}_t). \tag{2.1}$$

and

$$\boldsymbol{y}_t = h(\boldsymbol{x}_t, \phi, \boldsymbol{e}_t) \tag{2.2}$$

where $\boldsymbol{x}_t \in \mathbb{R}^m$ is an m-dimensional state vector, $\boldsymbol{y}_t \in \mathbb{R}^n$ is an n-dimensional observation (or measurement) vector, and $\boldsymbol{\eta}_t$ and $\boldsymbol{e}_t$ are serially uncorrelated state and observation noise processes, respectively.

The state vector, $\boldsymbol{x}_t$, also called the target state vector (Ristic et al. 2004) or hidden state process (Doucet et al. 2000), contains the true values of the processes of the system. The elements of this vector are unobserved and assumed to be generated by a first-order Markov process which implies that for $t = 1, 2, \ldots$, the conditional distribution of the state at time t given the states at all other times, $\boldsymbol{X}_{t-1} = \{\boldsymbol{x}_0, \boldsymbol{x}_1, \boldsymbol{x}_2, \ldots \boldsymbol{x}_{t-1}\}$, is equivalent to the conditional distribution of the state at time t given the state at time $t-1$ alone i.e. $p(\boldsymbol{x}_t)|\boldsymbol{X}_{t-1} = p(\boldsymbol{x}_t|\boldsymbol{x}_{t-1})$ (Agresti 2002). The distributions of the noise processes $\boldsymbol{\eta}$ and $\boldsymbol{e}$ are assumed known, and to be mutually independent and uncorrelated through time. The functions f and h along with their parameters θ and ϕ are assumed to be known, and possibly nonlinear, and the distribution of $\boldsymbol{x}_0$ is assumed known.

Equation (2.1) is known as the system equation, while equation (2.2) is known as the observation equation. The goal of the state space model then, is to produce the best estimate of the unobserved state vector, $\boldsymbol{x}_t$, from noisy observations, $\boldsymbol{y}_t$, taken at discrete time points by combining the system equation and the observation equation.

The state evolves according to a discrete-time stochastic model. The state space form of this state evolution equation is

$$\boldsymbol{x}_t = \boldsymbol{F}_t(\boldsymbol{x}_{t-1}, \theta) + \boldsymbol{\eta}_t \tag{2.3}$$

where $\boldsymbol{F}_t$ is an m-dimensional vector-valued function which relates the state at time t to the state at time $t-1$ via parameter vector θ, and $\boldsymbol{\eta}_t$ is an m-dimensional vector of

serially uncorrelated zero mean disturbances with covariance matrix, $\boldsymbol{\Sigma}_{\eta,t}$ with $E(\boldsymbol{\eta}_t) = 0$ and $\text{Cov}(\boldsymbol{\eta}_t) = \boldsymbol{\Sigma}_{\eta,t}$.

A multivariate time series of observations, $\boldsymbol{y}_t$, with additive noise, $\boldsymbol{e}_t$, are related to the state via the measurement equation which may be expressed as

$$\boldsymbol{y}_t = \boldsymbol{H}(\boldsymbol{x}_t, \phi) + \boldsymbol{e}_t \tag{2.4}$$

where $\boldsymbol{y}_t$ contains n-dimensional observation data, and $\boldsymbol{x}_t$ is an m-dimensional state vector, $\boldsymbol{H}$ is an $n \times m$ matrix which relates the observations to the unobserved state via parameter vector ϕ, and $\boldsymbol{e}_t$ is an n dimensional vector of serially uncorrelated zero mean disturbances with covariance matrix $\boldsymbol{\Sigma}_{e,t}$:

$$E(\boldsymbol{e}_t) = 0 \quad \text{and} \quad \text{Cov}(\boldsymbol{e}_t) = \boldsymbol{\Sigma}_{e,t} \ .$$

The error covariances for the observations $\boldsymbol{\Sigma}_{e,t}$ are based on knowledge about the accuracy of the measurements and the methodologies used to collect them. Only time homogeneous models are considered here, such that matrices $\boldsymbol{F}_t, \boldsymbol{\Sigma}_{\eta,t}, \boldsymbol{\Sigma}_{e,t}$, are independent of time, and may be written as $\boldsymbol{F}, \boldsymbol{\Sigma}_\eta, \boldsymbol{\Sigma}_e$.

An example of a slightly more complex state space model which will be used in this thesis is the multivariate second-order autoregressive process. Consider the AR(2) process expressed in state space form for times $1, 2, \ldots, t$,

$$\begin{aligned} \boldsymbol{x}_t &= \boldsymbol{A}_1 \boldsymbol{x}_{t-1} + \boldsymbol{A}_2 \boldsymbol{x}_{t-2} + \boldsymbol{\eta}_t \\ \boldsymbol{y}_t &= \boldsymbol{H}\boldsymbol{x}_t + \boldsymbol{e}_t \end{aligned} \tag{2.5}$$

where

$$\boldsymbol{x}_t = \begin{pmatrix} x_{t1} \\ x_{t2} \\ \vdots \\ x_{tk} \end{pmatrix}, \boldsymbol{y}_t = \begin{pmatrix} y_{t1} \\ y_{t2} \\ \vdots \\ y_{tk} \end{pmatrix}, \boldsymbol{A}_1 = \begin{pmatrix} a_{11} & 0 & \ldots & 0 \\ 0 & a_{12} & \ldots & \\ \vdots & & \ddots & \\ 0 & & & a_{1k} \end{pmatrix}, \boldsymbol{A}_2 = \begin{pmatrix} a_{21} & 0 & \ldots & 0 \\ 0 & a_{22} & \ldots & \\ \vdots & & \ddots & \\ 0 & & & a_{2k} \end{pmatrix},$$

$$\boldsymbol{\eta}_t = \begin{pmatrix} \eta_{t1} \\ \eta_{t2} \\ \vdots \\ \eta_{tk} \end{pmatrix} \text{ and } \boldsymbol{e}_t = \begin{pmatrix} e_{t1} \\ e_{t2} \\ \vdots \\ e_{tk} \end{pmatrix}.$$

A simple representation of the AR(2) process is if we have a state and data series, $\boldsymbol{X}, \boldsymbol{Y}$, with dimensions $m = 1$, and $n = 1$, matrix $\boldsymbol{H}$ as the $n \times m = 1 \times 1$ identity, 1, and matrices $\boldsymbol{A}_1, \boldsymbol{A}_2$ as $n \times n = 1 \times 1$ dimensional scalars (i.e. $\boldsymbol{A}_1 = a_1, \boldsymbol{A}_2 = a_2$), then (2.5) reduces to

$$\begin{aligned} x_t &= a_1 x_{t-1} + a_2 x_{t-2} + \eta_t \\ y_t &= x_t + e_t. \end{aligned}$$

Figure (2.1) is a realization of this AR(2) process that assigned $a_1 = 1.5$, $a_2 = -0.6$, $x_0 \sim \mathcal{N}(0, 1)$, with serially independent Gaussian noise, $\eta_t \sim \mathcal{N}(0, 0.21)$.

The advantage of the state space model approach is that the behaviour of several different components of a time series can be modelled separately, and the aggregate of these component models can be used to analyze the data. This makes state space models very general and able to handle a wide range of applications, from both the classical and Bayesian standpoints (Durbin and Koopman 2000).

Since models are never perfect, the state space framework is a way to combine observations from the system with the information provided by the system model in order to estimate important aspects of the system state as accurately as possible. The framework formally acknowledges that there is chance variation in the system model, and the error from this source is treated separately from the measurement (e.g. instrument) error. The state estimation procedure is known as *filtering*, and is reviewed in the following section.

2.2.1 Filtering

In the context of stochastic processes, a filter is a data processing algorithm which provides an estimator of the true state of a system based upon a set of observations related to the system (Maybeck 1979). Using the state space formulation of a model as in equations (2.1) and (2.2), the goal of filtering is to calculate best estimates (filtered estimates) of the state, $\boldsymbol{x}_t$, of the system equation using all of the available information.

From a Bayesian perspective, obtaining filtered estimates of $\boldsymbol{x}_t$ involves recursively updating the degree of belief in the state at time t given all available data $\boldsymbol{Y}_t$ up to time t, where $\boldsymbol{Y}_t = \{\boldsymbol{y}_1, \boldsymbol{y}_2, \ldots, \boldsymbol{y}_t\}$ is the set of all observations up to and including time t. For this, it is necessary either to calculate the posterior density analytically, or, if that is not possible, to construct an accurate approximation to the posterior density $p(\boldsymbol{x}_t|\boldsymbol{Y}_t, \theta)$. In general, the vector, θ, represents the set of parameters of the system equation (2.1) which control the evolution of the state, and is assumed known unless otherwise specified. The posterior

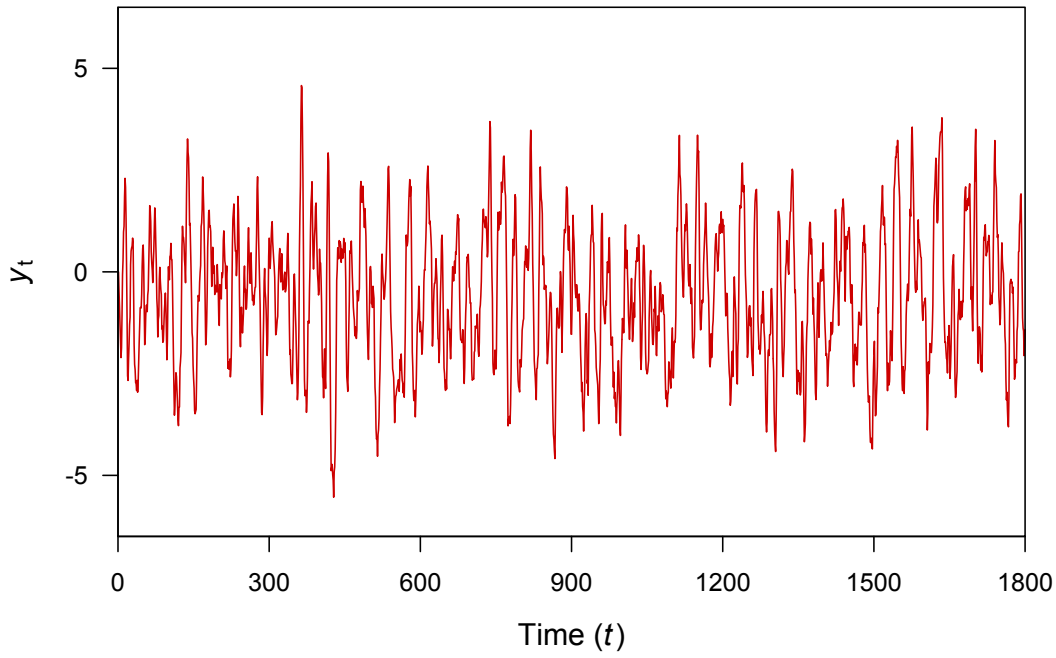

Figure 2.1: A realization of the AR(2) process with $a_1 = 1.5, a_2 = -0.6, \eta_t \sim \mathcal{N}(0, 0.21)$

density incorporates information about the state $\boldsymbol{x}_t$ which is contained in the observations, $\boldsymbol{Y}_t$, the initial prior density of the state, $p(\boldsymbol{x}_0)$, and the (assumed known) parameter vector, θ. Therefore, filtering is an estimation problem for which we are concerned with estimates of the current state obtained from an accurate approximation of the (posterior) density $p(\boldsymbol{x}_t|\boldsymbol{Y}_t, \theta)$. Below is an introduction to prediction of models in state space form, for further details see Kitagawa (1987) and Harvey (1989).

2.2.2 Filtering Solution

Assuming a dynamical model following (2.1, 2.2), with known state dynamics (2.3), $F(\boldsymbol{x}_t, \theta)$, and observation functions (2.4), $H(\boldsymbol{x}_t, \phi)$, with known initial prior density for the state vector, $p(\boldsymbol{x}_0)$, and known system parameter vector, θ, then filtering may be conducted to recursively obtain the posterior density at all time points via a two-step procedure.

The first stage is a prediction stage using system equation (2.3) and the second stage is an update stage that uses the observation equation (2.4). In the prediction step the system model is used to predict the state probability density function forward from one measurement

time to the next. The update step then uses the new data observation to modify the density function. The prediction step generally deforms and increases the variance of the state density due to the fact that the state is subject to unknown disturbances which are modeled as noise processes, $\boldsymbol{\eta}_t$. Generally, the update step tightens the prediction density using the latest observation available and Bayes theorem. Following Ristic et al. (2004), the prediction and update steps are as follows:

1. **Prediction step**
   Assuming the probability density of the state at time $t-1$, $p(\boldsymbol{x}_{t-1}|\boldsymbol{Y}_{t-1}, \theta)$ is available, the prediction step uses the system model (2.1) to calculate the *forecast* density of the state vector $\boldsymbol{x}_t$ at time t

$$p(\boldsymbol{x}_t|\boldsymbol{Y}_{t-1}, \theta) = \int p(\boldsymbol{x}_t|\boldsymbol{x}_{t-1}, \boldsymbol{Y}_{t-1}, \theta) p(\boldsymbol{x}_{t-1}|\boldsymbol{Y}_{t-1}, \theta) d\boldsymbol{x}_{t-1} \tag{2.6}$$

   Equation (2.3) is a Markov process of order one implying that $p(\boldsymbol{x}_t|\boldsymbol{x}_{t-1}, \boldsymbol{Y}_{t-1}, \theta) = p(\boldsymbol{x}_t|\boldsymbol{x}_{t-1}, \theta)$. Thus (2.6) becomes

$$p(\boldsymbol{x}_t|\boldsymbol{Y}_{t-1}, \theta) = \int p(\boldsymbol{x}_t|\boldsymbol{x}_{t-1}, \theta) p(\boldsymbol{x}_{t-1}|\boldsymbol{Y}_{t-1}, \theta) d\boldsymbol{x}_{t-1}, \tag{2.7}$$

   where the probability known as the transitional density $p(\boldsymbol{x}_t|\boldsymbol{x}_{t-1}, \theta)$, is defined in the process equation (2.3), and $p(\boldsymbol{x}_{t-1}|\boldsymbol{Y}_{t-1}, \theta)$ is the target or posterior probability density of the state at time $t-1$. The resulting forecast density $p(\boldsymbol{x}_t|\boldsymbol{Y}_{t-1}, \theta)$ gives the best prediction of the state at time t before the observation $\boldsymbol{y}_t$ is assimilated, and is interpreted in a Bayesian context as the prior for time step t.

2. **Update step**
   When an observation becomes available for time t, we update the forecast density (and our knowledge about the state) by updating the state through assimilation of the observation, $\boldsymbol{y}_t$, using Bayes' rule and assuming a Markov process

$$\begin{aligned} p(\boldsymbol{x}_t|\boldsymbol{Y}_t, \theta) &= p(\boldsymbol{x}_t|\boldsymbol{y}_t, \boldsymbol{Y}_{t-1}, \theta) \\ &= \frac{p(\boldsymbol{y}_t|\boldsymbol{x}_t, \boldsymbol{Y}_{t-1}, \theta) p(\boldsymbol{x}_t|\boldsymbol{Y}_{t-1}, \theta)}{p(\boldsymbol{y}_t|\boldsymbol{Y}_{t-1}, \theta)} \\ &= \frac{p(\boldsymbol{y}_t|\boldsymbol{x}_t, \theta) p(\boldsymbol{x}_t|\boldsymbol{Y}_{t-1}, \theta)}{p(\boldsymbol{y}_t|\boldsymbol{Y}_{t-1}, \theta)}. \end{aligned} \tag{2.8}$$

The denominator is a normalizing constant that can be written as

$$p(\boldsymbol{y}_t|\boldsymbol{Y}_{t-1},\theta) = \int p(\boldsymbol{y}_t|\boldsymbol{x}_t,\theta)p(\boldsymbol{x}_t|\boldsymbol{Y}_{t-1},\theta)d\boldsymbol{x}_t.$$

showing its dependence on the likelihood function, $p(\boldsymbol{y}_t|\boldsymbol{x}_t,\theta)$, defined by the measurement model (2.4), and the known parameters for the distribution of $\boldsymbol{e}_t$.

Thus, the posterior density $p(\boldsymbol{x}_t|\boldsymbol{Y}_t,\theta)$ is defined as a scaled product of the likelihood function of $\boldsymbol{y}_t$ determined from the measurement model in (2.4), $p(\boldsymbol{y}_t|\boldsymbol{x}_t,\theta)$, and the forecast density from the prediction step, $p(\boldsymbol{x}_t|\boldsymbol{Y}_{t-1},\theta)$. Equations (2.7) and (2.8) make up the recurrence relation that is the basis for the optimal Bayesian solution (Ristic et al. 2004). This recurrence relation allows for the calculation of the posterior density $p(\boldsymbol{x}_t|\boldsymbol{Y}_t,\theta)$, which in turn allows for calculation of state estimates via the Kalman filter which are optimal only in the restrictive case when the problem is limited to a linear system with Gaussian errors.

In the case where a state space model is nonlinear and/or non-Gaussian, there is no analytic solution to the filtering problem. Various different filtering strategies to the optimal solution have been proposed, all of which consist of analytic approximations, numerical approximations, and sampling approaches. The class of Monte Carlo filters called *particle filters* can handle both nonlinear models, and non-Gaussian errors. Specifically, the focus here will be on the Monte Carlo based sequential importance resampling filter suggested by Kitagawa (1996), detailed in Arulampalam et al. (2002) and Ristic et al. (2004), and described in the following section.

2.2.3 Sequential Monte Carlo Methods

A particle filter, also known as a sequential Monte Carlo method, is a sophisticated model estimation technique based on simulation. Particle filter methods approximate the entire posterior probability density function through time by a set of random samples, called particles, with associated weights. These particles naturally follow the movement of the state since they are potential realizations of the state of the system, generated from the system equation. The set of weighted particles can then be used to compute estimates of interest, such as the mean of the estimated posterior density. The particle filter presented below uses Monte Carlo methods, which rely on the idea of using samples to characterize entire probability distributions.

Monte Carlo Integration

Monte Carlo integration relies on the numerical evaluation of a multivariate integral, and forms the basis of sequential Monte Carlo (MC) estimation. The assumption is that $N \gg 1$ samples are drawn from the probability density, $p(\boldsymbol{x})$, of a random variable $\boldsymbol{x}$, $\{x^{(1)}, \ldots, x^{(N)}\}$, and can be used to determine the expected value of any function $f(\boldsymbol{x})$ provided that expectation exists.

Let the function $g(\boldsymbol{x})$ in the integral,

$$\mathcal{I} \quad = \quad \int g(\boldsymbol{x}) d\boldsymbol{x}, \tag{2.9}$$

be factored into $g(\boldsymbol{x}) = f(\boldsymbol{x})p(\boldsymbol{x})$ where $p(\boldsymbol{x})$ is a probability density function, and $f(\boldsymbol{x})$ is a function, then (2.9) can be rewritten

$$\mathcal{I} = \int f(\boldsymbol{x})p(\boldsymbol{x}) d\boldsymbol{x}. \tag{2.10}$$

Monte Carlo methods make the assumption that it is possible to draw samples distributed according to $p(\boldsymbol{x})$, for which the Monte Carlo estimate of (2.9) and (2.10) is the sample mean

$$\mathcal{I}_N \quad = \quad \frac{1}{N} \sum_{n=1}^{N} f(x^{(n)}) \ .$$

This sample mean is an unbiased and consistent estimator of $\mathcal{I}$ if the samples $\{x^{(1)}, \ldots, x^{(N)}\}$ are independent, and $\mathcal{I}_N$ will almost surely converge to $\mathcal{I}$ according to the law of large numbers. If the variance of $f(\boldsymbol{x})$ is finite then the central limit theorem holds, and the estimation error converges in distribution

$$\lim_{N \to \infty} \sqrt{N} \left(\mathcal{I}_N - \mathcal{I} \right) \sim \mathcal{N}(0, \sigma^2) \tag{2.11}$$

where σ^2 is the variance of $f(\boldsymbol{x})$. The error of this Monte Carlo estimate, is of order $O(N^{-1/2})$, which means that the rate of convergence is not dependent of the dimension of the integrand. This is an important feature of Monte Carlo integration methods since all deterministic numerical integration methods have rates of convergence which decrease as the dimension of the integrand increases (Lange 2010). This property makes Monte Carlo integration especially appealing for high-dimensional states.

In the Bayesian context, the density function $p(\boldsymbol{x})$ is the posterior density of interest in the filtering problem. Usually sampling directly from $p(\boldsymbol{x})$ is not possible due to its

multivariate and often non-standard properties. The next section describes a method of sampling a density similar to $p(\boldsymbol{x})$ of (2.10) that's easy to sample from to obtain Monte Carlo estimates of $\mathcal{I}$.

Importance Sampling

We would like to be able to sample directly from $p(\boldsymbol{x})$, but this is rarely possible. Importance sampling is a technique which allows Monte Carlo estimates of the integral $\int f(\boldsymbol{x})p(\boldsymbol{x})d\boldsymbol{x}$ to be made without sampling directly from $p(\boldsymbol{x})$. Instead we select a density $q(\boldsymbol{x})$ that is simple to sample from and is similar to $p(\boldsymbol{x})$. The samples drawn from $q(\boldsymbol{x})$ are then used to calculate $\mathcal{I}_N$, the estimate of $\mathcal{I}$. The probability density $q(\boldsymbol{x})$ is known as the importance density and is selected so that it has the same support as $p(\boldsymbol{x})$, that is,

$$q(\boldsymbol{x}) > 0 \text{ whenever } p(\boldsymbol{x}) > 0 \text{ for all } \boldsymbol{x} \in \mathbb{R}.$$

If this condition holds, then the integral $\mathcal{I}$ can be rewritten as

$$\mathcal{I} = \int f(\boldsymbol{x})p(\boldsymbol{x})d\boldsymbol{x} = \int f(\boldsymbol{x})\frac{p(\boldsymbol{x})}{q(\boldsymbol{x})}q(\boldsymbol{x})d\boldsymbol{x}. \tag{2.12}$$

An importance sampling estimator of (2.12) is computed by generating N independent sample points from the importance density, $q(\boldsymbol{x})$, that is as close to $p(\boldsymbol{x})$ as possible, and assigning each point an importance weight

$$\tilde{w}(\boldsymbol{x}) \propto \frac{p(\boldsymbol{x}^{(n)})}{q(\boldsymbol{x}^{(n)})}. \tag{2.13}$$

The importance weights are proportional to the relative probability that the sample was generated by the target density, $p(\boldsymbol{x})$, versus by the importance density, $q(\boldsymbol{x})$. In other words, particles which are close to the actual observation at a particular time step will be given weights closer to 1, and particles far from the actual observation will be given weights closer to 0. The model used to evaluate the weights depends on the observation equation (2.2), and the integral $\mathcal{I}$ can be estimated by the weighted sum

$$\mathcal{I}_N = \frac{1}{N}\sum_{n=1}^{N} f(\boldsymbol{x}^{(n)})\tilde{w}(\boldsymbol{x}^{(n)}).$$

If the normalizing factor of the desired density $p(\boldsymbol{x})$ is unknown, we normalize the importance weights and replace $\tilde{w}(\boldsymbol{x}^{(n)})$ with $w(\boldsymbol{x}^{(n)})$

$$w(\boldsymbol{x}^{(n)}) = \frac{\tilde{w}(\boldsymbol{x}^{(n)})}{\sum_{n=1}^{N}\tilde{w}(\boldsymbol{x}^{(n)})}.$$

In our setting, the target density $p(\boldsymbol{x})$ is the general filtering solution, or posterior density at time t, $p(\boldsymbol{x}_t|\boldsymbol{Y}_t, \theta)$ of (2.8). A common choice of importance density, $q(\boldsymbol{x})$, and the choice used in the following application because it is easy to generate from, is the transition density, $p(\boldsymbol{x}_t|x_{t-1}, \theta)$, obtained from the system equation (2.3) during the predictive step.

Sequential Importance Sampling

The main objective of Sequential Importance Sampling (SIS) is to represent the posterior density function by a set of random sample particles (called support points) that have associated weights. The conceptual idea is to apply the importance sampling technique sequentially to the filtering solution. By sampling $N \gg 1$ particles, various estimates can be computed for which it can be shown that the SIS estimates approach the optimal Bayesian estimators as $N \to \infty$.

Suppose that $\boldsymbol{X}_t = \{\boldsymbol{x}_0, \boldsymbol{x}_1, \ldots, \boldsymbol{x}_t\}$ denotes a sequence of states up to time t. The joint posterior distribution of the states at time t is denoted as $p(\boldsymbol{X}_t|\boldsymbol{Y}_t)$ and the marginal distribution of the state at time t as $p(\boldsymbol{x}_t|\boldsymbol{Y}_t)$. Suppose the set of support points, $\{\boldsymbol{X}_t^{(n)}; n = 1, \ldots, N\}$ with associated normalized weights $\{w_t^{(n)}; n = 1, \ldots, N\}$ and $\sum_{n=1}^{N} w_t^{(n)} = 1$ characterize the joint posterior density. The joint posterior distribution of the state at time t, $p(\boldsymbol{X}_t|\boldsymbol{Y}_t)$, is then approximated by

$$p(\boldsymbol{X}_t|\boldsymbol{Y}_t) \approx \sum_{n=1}^{N} w_t^{(n)} \delta(\boldsymbol{X}_t - \boldsymbol{X}_t^{(n)}) \tag{2.14}$$

where $\delta(\cdot)$ represents the idealized Dirac delta measure with the properties

1. $\delta(\boldsymbol{X}_t - \boldsymbol{X}_t^{(n)}) = 0$ for $\boldsymbol{X}_t \neq \boldsymbol{X}_t^{(n)}$,
2. $\delta(\boldsymbol{X}_t - \boldsymbol{X}_t^{(n)}) = \infty$ for $\boldsymbol{X}_t = \boldsymbol{X}_t^{(n)}$, and
3. $\int_{-\infty}^{\infty} \delta(\boldsymbol{X}_t - \boldsymbol{X}_t^{(n)})\, dx = 1$,

and can be thought of as an infinitely high, infinitely thin spike at the origin, with the total area under the spike equal to 1, which conceptually represents an idealized point mass. Equation (2.14) is a discrete weighted approximation of the true joint posterior distribution $p(\boldsymbol{X}_t|\boldsymbol{Y}_t)$. If the ensemble of particles $\{\boldsymbol{X}_t^{(n)}; n = 1, \ldots, N\}$ were taken from the importance density $q(\boldsymbol{X}_t|\boldsymbol{Y}_t)$ then the weights of equation (2.14) are defined as in (2.13) as

$$w_t^{(n)} = \frac{p(\boldsymbol{X}_t^{(n)}|\boldsymbol{Y}_t)}{q(\boldsymbol{X}_t^{(n)}|\boldsymbol{Y}_t)}. \tag{2.15}$$

Therefore, these sequential MC particle filter methods approximate the predicted distribution by a set of Dirac delta functions, referred to as an ensemble of particles $\{\boldsymbol{X}_t^{(n)}\}$ corresponding to the importance weights $w_t^{(n)}$.

Suppose an ensemble of particles is available that approximates the joint density at time $t-1$, $p(\boldsymbol{X}_{t-1}|\boldsymbol{Y}_{t-1})$. When the next data point $\boldsymbol{y}_t$ is observed at time t, our aim is to approximate the joint density $p(\boldsymbol{X}_t|\boldsymbol{Y}_t)$. Selecting the importance density such that it can be factored into the product of the marginal distribution of the old state and the joint distribution of the new observation

$$q(\boldsymbol{X}_t|\boldsymbol{Y}_t) = q(\boldsymbol{x}_t|\boldsymbol{Y}_t, \boldsymbol{X}_{t-1})q(\boldsymbol{X}_{t-1}|\boldsymbol{Y}_{t-1}) \tag{2.16}$$

then samples $\boldsymbol{X}_t^{(n)} \sim q(\boldsymbol{X}_t|\boldsymbol{Y}_t)$ can be obtained by augmenting the existing samples $\boldsymbol{X}_{t-1}^{(n)} \sim q(\boldsymbol{X}_{t-1}|\boldsymbol{Y}_{t-1})$ with the new state $\boldsymbol{x}_t^{(n)} \sim q(\boldsymbol{x}_t|\boldsymbol{X}_{t-1}, \boldsymbol{Y}_t)$. The updated weights can be determined by expanding the probability $p(\boldsymbol{X}_t|\boldsymbol{Y}_t)$ using Bayes rule such that

$$\begin{aligned} p(\boldsymbol{X}_t|\boldsymbol{Y}_t) &= \frac{p(\boldsymbol{y}_t|\boldsymbol{X}_t, \boldsymbol{Y}_{t-1})\ p(\boldsymbol{X}_t|\boldsymbol{Y}_{t-1})}{p(\boldsymbol{y}_t|\boldsymbol{Y}_{t-1})} \\ &= \frac{p(\boldsymbol{y}_t|\boldsymbol{X}_t, \boldsymbol{Y}_{t-1})\ p(\boldsymbol{x}_t|\boldsymbol{X}_{t-1})p(\boldsymbol{X}_{t-1}|\boldsymbol{Y}_{t-1})}{p(\boldsymbol{y}_t|\boldsymbol{Y}_{t-1})} \\ &\propto p(\boldsymbol{y}_t|\boldsymbol{X}_t, \boldsymbol{Y}_{t-1})p(\boldsymbol{x}_t|\boldsymbol{X}_{t-1})\ p(\boldsymbol{X}_{t-1}|\boldsymbol{Y}_{t-1}) \end{aligned} \tag{2.17}$$

By substituting the importance density (2.16) and (2.17) into (2.15), the weight update equation becomes

$$\begin{aligned} w_t^{(n)} &\propto \frac{p(\boldsymbol{y}_t|\boldsymbol{x}_t^{(n)})p(\boldsymbol{x}_t^{(n)}|\boldsymbol{x}_{t-1}^{(n)})p(\boldsymbol{X}_{t-1}^{(n)}|\boldsymbol{Y}_{t-1}^{(n)})}{q(\boldsymbol{x}_t^{(n)}|\boldsymbol{Y}_t, \boldsymbol{X}_{t-1}^{(n)})q(\boldsymbol{X}_{t-1}^{(n)}|\boldsymbol{Y}_{t-1})} \\ &\propto \frac{p(\boldsymbol{y}_t|\boldsymbol{x}_t^{(n)})p(\boldsymbol{x}_t^{(n)}|\boldsymbol{x}_{t-1}^{(n)})}{q(\boldsymbol{x}_t^{(n)}|\boldsymbol{Y}_t, \boldsymbol{X}_{t-1}^{(n)})}\ w_{t-1}^{(n)} \end{aligned} \tag{2.18}$$

If the importance density $q(\boldsymbol{x}_t^{(n)}|\boldsymbol{X}_{t-1}^{(n)}, \boldsymbol{Y}_t)$ of (2.18) is dependent only on the state at time $t-1$, $\boldsymbol{x}_{t-1}^{(n)}$, rather than the entire path of the state, $\boldsymbol{X}_{t-1}^{(n)}$, and the new observation received at time t, $\boldsymbol{y}_t$, rather than all previous observed data, $\boldsymbol{Y}_{t-1}^{(n)}$, then $q(\boldsymbol{x}_t^{(n)}|\boldsymbol{X}_{t-1}^{(n)}, \boldsymbol{Y}_t) = q(\boldsymbol{x}_t^{(n)}|\boldsymbol{x}_{t-1}^{(n)}, \boldsymbol{y}_t)$. When this equality is true then the estimate of the marginal posterior distribution, $p(\boldsymbol{x}_t|\boldsymbol{Y}_t)$, can be obtained by use of a modified weight that is a function of

the weights at the previous time step, $w_{t-1}^{(n)}$, the likelihood, $p(\boldsymbol{y}_t|\boldsymbol{x}_t^{(n)})$, and the transition density, $p(\boldsymbol{x}_t^{(n)}|\boldsymbol{x}_{t-1}^{(n)})$. The modified weights are estimated as

$$w_t^{(n)} \propto w_{t-1}^{(n)} \frac{p(\boldsymbol{y}_t|\boldsymbol{x}_t^{(n)})p(\boldsymbol{x}_t^{(n)}|\boldsymbol{x}_{t-1}^{(n)})}{q(\boldsymbol{x}_t^{(n)}|\boldsymbol{y}_t, \boldsymbol{x}_{t-1}^{(n)})} \tag{2.19}$$

and the posterior filtered distribution may then be approximated, as in (2.14), by

$$p(\boldsymbol{x}_t|\boldsymbol{Y}_t) \approx \sum_{n=1}^{N} w_t^{(n)} \delta(\boldsymbol{x}_t - \boldsymbol{x}_t^{(n)}). \tag{2.20}$$

Thus, only $\boldsymbol{x}_t^{(n)}$ need be stored to solve for the filtered solution, and all previous state estimates $\boldsymbol{X}_{t-1}^{(n)}$, and previous observed data $\boldsymbol{Y}_{t-1}^{(n)}$ can be discarded. It can be shown (e.g. Arulampalam et al. 2002) that as $N \to \infty$ the approximation (2.20) approaches the true posterior density $p(\boldsymbol{x}_t|\boldsymbol{Y}_t)$. Thus, no explicit assumption about the form of the target (posterior) density $p(\boldsymbol{x}_t|\boldsymbol{Y}_t)$ (2.20) is made, and the SIS algorithm thus consists of recursive propagation of the weights and sample particles as each measurement is received sequentially.

The major limitation of SIS is impoverishment of the particle ensemble due to the problem of a limited number of particles sharing the majority of the importance weight. As this is nearly impossible to avoid in SIS, one solution to this lack of diversity in the particle ensemble is the use of resampling.

Sequential Importance Resampling

The Sequential Importance Resampling (SIR) filter is derived from the SIS algorithm by performing a resampling step at every time step to conserve the same number of particles at each time step. The algorithm involves resampling with replacement the set of weighted particles $\{\boldsymbol{x}_t^{(n)}\}$ with probabilities equal to the weights of (2.19). By resampling with replacement $N \gg 1$ times from the (discrete approximation of the) posterior density (2.20), new samples $\{\boldsymbol{x}^{(n')}; n = 1, \ldots, N\}$ that are i.i.d. are generated, now each with equal weights $w_t^{(n')} = 1/N$. The SIR algorithm recursively evaluates the importance weights for each sample, and resamples and propagates particles to avoid degeneracy of the ensemble at each time step. Sequential MC methods such as SIR are also known as *particle filtering* as each time step is an MC realization in which each particle's trajectory through the state space can be mapped. Each particle filter step prunes and expands particles in a way analogous to Darwinian selection.

A generic SIR algorithm is summarized in Figure (2.2), and a more sophisticated SIR algorithm that includes parameter estimation as well as state estimation is discussed in detail in Section (2.3).

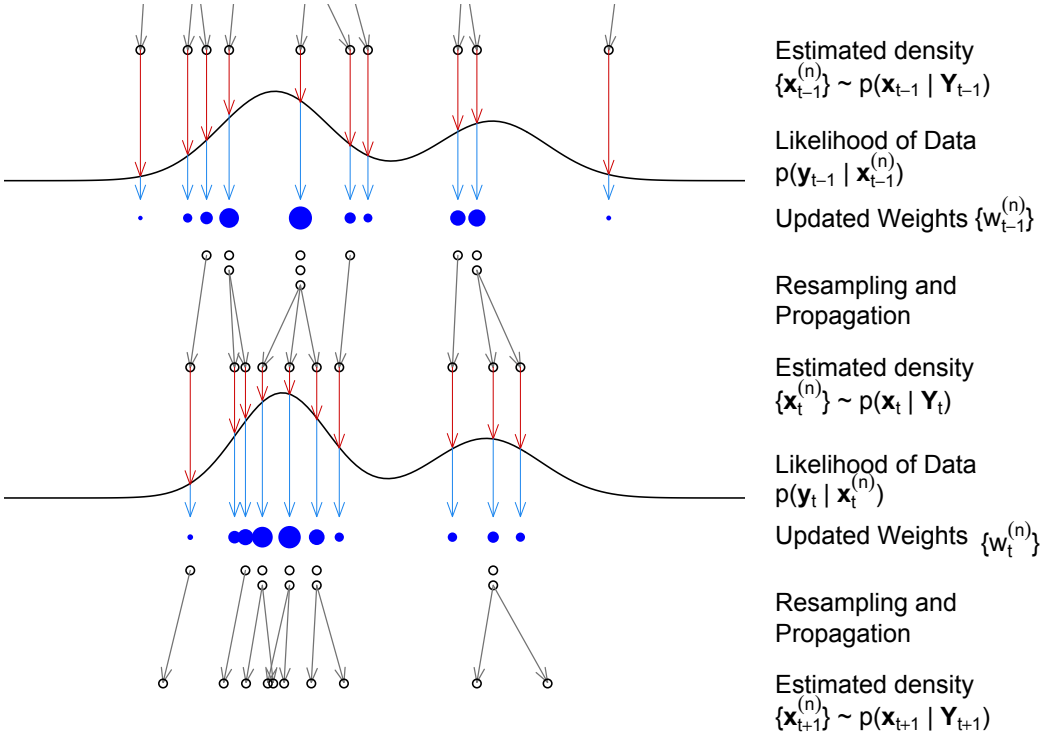

Figure 2.2: Two iterations of a generic SIR particle filter for $N = 10$ particles. The new state vector is simulated for each particle using the process equation (Equation 2.1). The observation update (Equation 2.2) is achieved by updating the weights with a likelihood function following (Equation 2.19). We then apply importance resampling to the weighted particles using sampling with replacement. After this resampling step, all particles have uniform weights $w^n = 1/N$. This ensemble of 10 particles $\boldsymbol{x}_t^{(n)}; n = 1, \ldots, 10$ represents our posterior estimate of the state at time t. Repeat. (Adapted from Ristic et al. 2004).

Particle filter methods can also be studied by extending the state $\boldsymbol{x}_t$ to include a parameter vector, which will be discussed in Section (2.3). The implementation of the SIR particle filter in the problem of estimating time-varying parameters, is a non-standard application and will be discussed in Section (2.3.1) below.

2.3 Parameter Estimation Using Particle Filters

Up until now, we have assumed that the model dynamics and the parameters that describe this process are known without error. Although this may be reasonable in an engineering application, it is seldom the case in ecology that we know the "*true*" parameters that determine the dynamics of a given process. In such applications where the parameters as well as the state need to be estimated simultaneously, the state vector can be reconfigured by appending the parameter vector to the state vector and both components then form part of the "prior" assumptions (or forecast density). The methods described next show that the final prediction of the target density, $p(\boldsymbol{x}_t|\boldsymbol{Y}_t, \theta)$, allows, as before, for uncertainties in the data and the process model, but also includes estimates and uncertainties in parameters.

To estimate parameters in a sequential MC state space model, two main approaches are available. The first approach is to compute a likelihood function using a particle filter. However, natural Monte Carlo variation means the likelihood surface is rough and difficult to maximize (Doucet and Tadić 2003), and it is computationally inefficient except where closed form estimators exist as in linear, Gaussian problems (Kitagawa 1998, Hürzeler and Künsch 2000). A second approach to parameter estimation that works for both linear and nonlinear systems, and non-Gaussian state space models, was developed by Kitagawa (1998). To estimate the unknown parameters of the system, Kitagawa (1998) considered a Bayesian estimation procedure to approximate the marginal posterior densities of both the state and the unknown parameters simultaneously by appending the parameter vector θ to the state vector. The ability of the state space model to include an augmented state increases its applicability to a wider range of problems including those for which parameters in the system equation are not well understood, and can't be assumed to be known.

Consider first the situation in which time-invariant parameters of a system are to be estimated. For all $t > 0$, the filter sample, $\theta_t^{(n)}$, is a sub-sample of the previous posterior density sample, $\theta_{t-1}^{(n)}$, which will only have some values in the region where the posterior distribution is concentrated, similar to the sample degeneracy of the SIS algorithm. By adding

a small amount of noise to the parameter transitions, the amount of sample degeneracy may be reduced. We denote the artificial evolution of the parameter vector θ with perturbation ν_t by the following transition equation

$$\theta_t = \theta_{t-1} + \nu_t. \tag{2.21}$$

Note that unlike the state $\boldsymbol{x}_t$, the parameter vectors θ_{t-1}, and θ_t in Equation (2.21) are not time-dependent, but rather the subscripts track the value of the estimate of θ at time $t-1$ and t.

We write the augmented state vector $\tilde{\boldsymbol{x}}_t$ by augmenting the state vector $\boldsymbol{x}_t$ to include the unknown parameter vector θ_t

$$\tilde{\boldsymbol{x}}_t = \begin{pmatrix} \boldsymbol{x}_t \\ \theta_t \end{pmatrix},$$

and by making the parameters in the process model (2.3) follow a random walk (i.e. $\theta_t = \theta_{t-1} + \nu_t$), the augmented state evolution equation can be written as follows

$$\begin{pmatrix} \boldsymbol{x}_t \\ \theta_t \end{pmatrix} = \begin{pmatrix} f(\boldsymbol{x}_{t-1}, \theta_t) \\ \theta_{t-1} \end{pmatrix} + \begin{pmatrix} \boldsymbol{\eta}_t \\ \nu_t \end{pmatrix} \tag{2.22}$$

or

$$\tilde{\boldsymbol{x}}_t = g(\tilde{\boldsymbol{x}}_{t-1}) + \tilde{\boldsymbol{\eta}}_t. \tag{2.23}$$

Here, the augmented system noise $\tilde{\boldsymbol{\eta}}_t$ now includes both the system noise, $\boldsymbol{\eta}_t$, as well as the disturbance term, ν_t, for the random walk of θ. The operator g includes both the system dynamics f and the random walk for the parameters θ to be estimated. The observation equation (2.2) is also augmented and is written as

$$\boldsymbol{y}_t = (H \quad \mathbf{0})\ \tilde{\boldsymbol{x}}_t,$$

where $\mathbf{0}$ is a vector of zeros to imply that we cannot observe the parameters, θ, directly. The augmented state space model can then be solved using the same sequential Monte Carlo methods as for regular state space models, since Equation (2.23) is a reformulation of the state equation (2.3).

For the estimation of a (slowly) time-varying parameter the augmented state can be redefined to incorporate an evolving parameter, however this is more complex involving multiple passes of the filter, but is also based on sequential importance resampling methods similar to those discussed in Section (2.2.3). Next, we describe the algorithm for the augmented state estimation problem of a static parameter vector θ using a particle filter.

2.3.1 Augmented State Space Solution via Particle Filtering

The goal of the augmented state space model is to produce estimates of both the state and the parameters of the system equation at each time step.

Consider a sample from an augmented state vector at time $t-1$ in which the usual state vector has been augmented to include the unknown parameter vector θ. Each particle of the sample has the same dimension as the augmented state vector, $\tilde{\boldsymbol{x}}_{t-1} \in \mathbb{R}^m$. The particles are each denoted by $\tilde{\boldsymbol{x}}_{t-1}^{(n)}$ where $n = 1, \ldots, N$ and the complete sample ensemble of N members is designated as $\{\tilde{\boldsymbol{x}}_{t-1}^{(n)}\}$.

Suppose we are at time $t-1$, and the sample is drawn from the (multivariate) probability density function $p(\tilde{\boldsymbol{x}}_{t-1}|\boldsymbol{Y}_{t-1})$, where

$$\{\tilde{\boldsymbol{x}}_{t-1}^{(n)}\} \sim p(\tilde{\boldsymbol{x}}_{t-1}|\boldsymbol{Y}_{t-1}), \qquad n = 1, \ldots, N, \tag{2.24}$$

then this provides a sample-based approximation to the posterior density of the augmented state. As before, the particle filtering algorithm is a recursive, Markov process, so that we need only consider the time transition of the system from $t-1$ to t to specify the procedure completely. The target distribution of the particle filter at time t is the marginal probability distribution conditional on having observations up to time t as represented by

$$\{\tilde{\boldsymbol{x}}_{t}^{(n)}\} \sim p(\tilde{\boldsymbol{x}}_{t}|\boldsymbol{Y}_{t}), \qquad n = 1, \ldots, N. \tag{2.25}$$

Thus, SIR is a sample-based method to carry out the transition from the augmented state solution at time $t-1$, Equation (2.24), to that at time t, Equation (2.25). The generic SIR state space algorithm of Figure (2.2) also provides a pictorial representation of the SIR augmented state space algorithm by replacing $\boldsymbol{x}$ with $\tilde{\boldsymbol{x}}$. A detailed algorithm for the particle filter used to generate a sample from the target posterior of the augmented state vector $\{\tilde{\boldsymbol{x}}_{t}^{(n)}\}$ from the current sample, $\{\tilde{\boldsymbol{x}}_{t-1}^{(n)}\}$ is as follows:

1. **Prediction Step**

   The prediction step uses the system model (2.1, 2.3) to predict the augmented state forward from time $t-1$ (2.24) to time t (2.25). The state space form (2.3) is now defined as $\tilde{\boldsymbol{x}}_t = \boldsymbol{f}(\boldsymbol{x}_t, \theta_t) + \tilde{\boldsymbol{\eta}}_{\boldsymbol{t}}$, where the artificial evolution of the parameter vector, θ_t, includes a small perturbation as in (2.21), and the augmented system noise includes ν_t as in (2.22). We produce a sample from the forecast density, $p(\tilde{\boldsymbol{x}}_t|\boldsymbol{Y}_{t-1})$, designated

as $\{\tilde{\boldsymbol{x}}_{t|t-1}^{(n)}\}$ by treating each member of the sample at time $t-1$, $\{\tilde{\boldsymbol{x}}_{t-1}^{(n)}\}$ as an initial condition for prediction, i.e.

$$\tilde{\boldsymbol{x}}_t^{(n)} = f(\tilde{\boldsymbol{x}}_{t-1}^{(n)}) + \tilde{\boldsymbol{\eta}}_t^{(n)}, \qquad n = 1, \ldots, N. \tag{2.26}$$

2. **Update/Resampling/Filtering Step**
   We update the forecast density by drawing a sample with replacement from the forecast ensemble defined in the prediction step with a probability proportional to a set of importance weights. These weights are computed for each sample member (or particle) as in Equation (2.19), i.e.

$$w_t^{(n)} \quad \propto \quad w_{t-1}^{(n)} \, \frac{p(\boldsymbol{y}_t|\tilde{\boldsymbol{x}}_t^{(n)}) p(\tilde{\boldsymbol{x}}_t^{(n)}|\tilde{\boldsymbol{x}}_{t-1}^{(n)})}{q(\tilde{\boldsymbol{x}}_t^{(n)}|\boldsymbol{y}_t, \tilde{\boldsymbol{x}}_{t-1}^{(n)})} \quad = \quad p(\boldsymbol{y}_t|\tilde{\boldsymbol{x}}_{t|t-1}^{(n)}). \tag{2.27}$$

   If we select the importance density, $q(\cdot)$, to be the transition density $p(\tilde{\boldsymbol{x}}_t|\tilde{\boldsymbol{x}}_{t-1})$ obtained as a component of the forecast density of the prediction step, this simplifies the importance weights to be proportional to the likelihood $p(\boldsymbol{y}_t|\tilde{\boldsymbol{x}}_t^{(n)})$. Here, $p(\boldsymbol{y}_t|\tilde{\boldsymbol{x}}_{t|t-1}^{(n)})$ is the likelihood of the observations $\boldsymbol{y}_t$ conditional on knowledge of the predicted state $\boldsymbol{x}_{t|t-1}^{(n)} \sim p(\tilde{\boldsymbol{x}}_t^{(n)}|\tilde{\boldsymbol{x}}_{t-1}^{(n)})$ for the n^{th} sample member of the previous step. A standard weighted resampling with replacement of the sample $\{\tilde{\boldsymbol{x}}_{t|t-1}^{(n)}\}$ is undertaken using the corresponding weights $\{w_t^{(n)}\}$. After resampling, this yields a sample $\{\boldsymbol{x}_t^{(n)}\}$ from the target posterior density $p(\tilde{\boldsymbol{x}}_t|\boldsymbol{Y}_t)$.

Therefore given a starting value for the state at $t = 0$, this sequential importance resampling algorithm can be run forward in time to sequentially generate the required samples for estimation of the state $\boldsymbol{X}$ and associated static parameter vector θ for the duration of the time series. Often, however, it is limiting to have to assume that the true parameter vector θ of the system model (2.1) is time invariant, the following section discusses an application where we allow the parameter vector to (slowly) evolve over time.

2.3.2 Iterated Filtering

Non-stationary time series models in which the system parameters evolve slowly over time compared to the dynamics represent a non-standard application of the SIR particle filter. Identification of the appropriate time scale for this estimation is a key feature, and will depend on the system in which the analysis is being applied. Standard application of the

augmented state space model (2.23) with a parameter random walk having a constant variance, σ_ν^2, means that the estimated parameter vector is assumed static while following short period fluctuations in the time series data (Kitagawa 1998). This is not always desirable for analysis and inference. For example, it is often the case in economic, oceanographic and biological data, that strings of data collected over time are collected at a higher density than that of the evolution of the system. For example, short-term fluctuations in stock prices do not represent important shifts in the market, fine-scale changes in polar ice coverage do not represent changes in global climate, and velocity measurements at two-second intervals for a marine mammal do not represent important shifts in their behaviour. Hence, we reformulate the parameter estimation problem to focus on estimation at time scales that can be directly linked to more meaningful time frames of inference.

The analysis is sequentially carried out on short, overlapping segments (windows) of the observed time series. Within each data segment, fixed values for the parameters of the process model are estimated while the state estimates follow the time resolution of the time series data. By considering the time sequence of these static parameter estimates over the duration of the data record, the time variation of the movement parameters on the appropriate time scale for inference can be determined. This idea of evolving parameters is well established for non-stationary statistics which vary over time (Priestley 2004, Ch. 11), and has the advantage that the investigator can control the time scale for parameter estimation by choosing an appropriate data segment length.

The estimation procedure within an analysis time window also relies on the state augmentation procedures outlined above. The main idea is that by successively reducing the variance of the disturbance term, σ_ν^2, in the parameter random walk, the parameter vector, θ_t, ends up being fixed at a particular value. One way this can be accomplished is by using a single pass of a particle filter over an analysis time window and reducing σ_ν^2 with increasing t, until σ_ν^2 reaches a small value at the end of the analysis segment and the parameters, θ, are locked in at their estimated values (Kitagawa 1998, Liu and West 2001). As a related alternative, Ionides et al. (2006) suggest using multiple passes of a state-augmented particle filter, where each pass has a successively smaller σ_ν^2 and yields the maximum likelihood value for θ. We make use of a hybrid approach, using the multiple iterated filtering approach of Ionides et al. (2006) algorithm, but using unweighted averages at each pass of the particle filter and a simple ramp-down of the variance, σ_ν^2. This is done within a time windowing

sequential procedure and permits slowly time-varying parameters to be estimated over successive windows. The goal is to interpret these parameters as they slowly vary in time at a scale that is sensitive to changes at a broader time scale of relevance to the phenomenon under investigation.

2.3.3 Conclusion

We apply this iterated filtering method to the biological interpretation of marine mammal movement data collected at a resolution of 2 seconds, over a 24 hour period to infer behaviour. We discuss the details of the algorithm in the context of the analysis of these data in the following chapter.

2.4 Literature Cited

Agresti A. 2002. Categorical data analysis. Wiley-Interscience, 2nd edition.

Arulampalam, M., S. Maskell, N. Gordon, and T. Clapp. 2002. A tutorial on particle filters for online nonlinear/non-Gaussian Bayesian tracking. IEEE Transactions on Signal Processing, 50(2): 174-188.

Doucet, A., S. Godsill, and C. Andrieu. 2000. On sequential Monte Carlo sampling methods for Bayesian filtering. Statistics and Computing, 10: 197-208.

Doucet, A., V.B. Tadić. 2003. Parameter estimation in general state space models using particle methods. Annals of the Institute of Statistical Mathematics. 55(2): 406-422.

Durbin, J., and S.J. Koopman. 2000. Time series analysis of non-Gaussian observations based on state space models from both Classical and Bayesian perspectives. Journal of the Royal Statistical Society. Series B (Statistical Methodology), 62(1): 3-56.

Harvey, A.C. 1989. Forecasting, structural time series models and the Kalman filter. Cambridge University Press.

Hürzeler, M., and H. Künsch. 2000. Approximating and maximizing the likelihood for a general state-space model. In A. Doucet, N. de Freitas, and N. Gordon, editors, Sequential Monte Carlo Methods in Practice. New York. Springer-Verlag, New York.

Ionides, E.L., and C. Breto, and A.A. King. 2006. Inference for nonlinear dynamical systems. Proceedings of the National Academy of Sciences, 103: 18438-18443.

Kalman, R.E. 1960. A new approach to linear filtering and prediction problems. Transactions of the ASME, Journal of Basic Engineering, 82: 35-45.

Lange, K. 2010. Numerical Analysis for Statisticians, 2^{nd} Edition. Springer; New York, 600pp.

Liu, J., and R. Chen. 1998. Sequential Monte Carlo methods for dynamical systems. Journal of the American Statistical Association, 93: 1032-1044.

Liu, J., and M. West. 2001. Combined parameter and state estimation in simulation based filtering. In Sequential Monte Carlo Methods in Practice, Doucet, A., N. de Freitas, and N. Gordon (eds); Springer: New York; 197-217.

Kitagawa, G. 1987. Non-Gaussian state-space modeling of nonstationary time series. Journal of the American Statistical Association, 82: 1032-1063.

Kitagawa, G. 1996. Monte Carlo filter and smoother for non-Gaussian nonlinear state space models. Journal of Computational and Graphical Statistics, 5: 1-25.

Kitagawa, G. 1998. A self-organising state-space model. Journal of the American Statistical Association, 93: 1203-1215.

Maars, A.D. 2000. In-situ ellipsometry solutions using sequential Monte Carlo. In A. Doucet, N. de Freitas, and N. Gordon, editors, Sequential Monte Carlo Methods in Practice. New York. Springer-Verlag, New York.

Maybeck, P.S. 1979. Stochastic models, estimation, and control, volume 141 of Mathematics in Science and Engineering. Academic Press.

Pitt, M., and N. Shepard. 1999. Filtering via simulation: Auxiliary particle filters. Journal of the American Statistical Association, 94: 590-599.

Priestley, M.B. 2004. Spectral Analysis and Time Series. Academic Press. London. 890pp.

Rabiner, L.R. 1989. A tutorial on hidden Markov models and selected applications in speech recognition. Proceedings of the IEEE, 77: 257-286.

Ristic, B., S. Arulampalam, and N. Gordon. 2004. Beyond the Kalman Filter: Particle Filters for Tracking Applications. Artech House: Boston, 318pp.

Chapter 3

An application of the state-augmented particle filter for estimating animal movement parameters

3.1 An Application using a State-Augmented Particle Filter in the Estimation of Animal Movement Parameters

3.1.1 Introduction

Fine-scale archival, data-logging technology has given rise to a rapidly growing body of information on the movement of many marine animal species. It is widely recognized that these technological developments have far outpaced the analysis methods available for extracting meaningful biological information from these high-resolution and complex data types (Schick et al. 2008). To date, many applications of animal movement data have focused on reconstructing tracks from sparse and noisy fixes of geographical position, but as positional information improves and motion sensors are incorporated (Ropert-Coudert and Wilson 2005), there is strong interest in inferring behaviour from data with much higher time resolution (Polansky et al. 2010). Current statistical approaches often require ad-hoc processing to circumvent the inherent time dependence in this kind of data set (e.g. Bodkin

et al. 2007). However, these data are fundamentally time series whose salient character is their autocorrelated structure and the non-stationarity of variance. State space models provide a flexible framework for a unified treatment of tagging data and animal movement models and are recognized as a promising way forward (Patterson et al. 2008). Here, we propose a modeling approach that uses high-resolution movement data to estimate continuously varying behavioural parameters of movement models on the appropriate time scale.

Animal movement models are formulated as stochastic differential (or difference) equations, such as a correlated random walk (Morales et al. 2004, Codling et al. 2008). These movement equations can describe a wide variety of movement patterns depending on their parameter values. Our main premise is that, as the movement observations change over the course of the record, the parameters of the animal movement model embody the behavioural information and can be estimated directly using a state space model formulation. One possibility is to use a state space behavioural switching model (Jonsen et al. 2007, Patterson et al. 2009), but this requires a pre-definition of a small set of behavioural modes, and the probability of transitioning between them. We suggest and explore a different approach that directly estimates the time-varying parameters of a movement model with a state space model, and uses these estimates as a basis for behavioural inference.

We develop and illustrate the approach using a high resolution data set from an at-sea foraging track of a northern fur seal (*Callorhinus ursinus*). The analysis uses vertical velocity (dive) data recorded at two second time intervals, which are of a much higher resolution (and volume) than data used for other state space models that infer behaviour using only episodic horizontal position fixes (e.g. Jonsen et al. 2007). Moreover, these data serve to illustrate the non-trivial issues that arise in model selection, and the setup of a state space model for a practical example. We emphasize, however, that the approach is a general one, and applicable to most horizontal and/or vertical movement models and high resolution tag data.

3.1.2 Methods

The approach taken fits a stochastic animal movement model to noisy position or velocity data using a state space model, simultaneously estimating its parameters through a state augmentation procedure. The first part of a state space model is the state evolution equation, or movement model, which describes the movement process, and allows it to evolve

forward in time as (2.3),

$$\boldsymbol{x}_t = \boldsymbol{f}(\boldsymbol{x}_{t-1}, \theta_t) + \boldsymbol{n}_t \,,$$

where $\boldsymbol{x}_t$ is a column vector that represents the system state (either a position or velocity) at time t. The movement model is embodied in the operator $\boldsymbol{f}$ which depends on the previous state, $\boldsymbol{x}_{t-1}$, and a vector of movement parameters, $\boldsymbol{\theta}_t$, that can change through time. The system noise, $\boldsymbol{n}_t$, is assumed additive.

The second part of the state space model is the observation equation in which the observations are assimilated as in (2.4),

$$\boldsymbol{y}_t = \boldsymbol{H}\ \boldsymbol{x}_t + \boldsymbol{e}_t \tag{3.1}$$

where $\boldsymbol{y}_t$ is an observation of vertical velocity at time t. It is related to the state, $\boldsymbol{x}_t$, through the matrix $\boldsymbol{H}$ which allows for conversion between measured and modelled variables (say, observing position and modelling velocity), or partial observation of the state. The observation error process, $\boldsymbol{e}_t$, (again assumed additive) follows a specified probability distribution. More general nonlinear forms of (2.2) are possible, and parameters may also be associated with $\boldsymbol{e}_t$ or $\boldsymbol{H}$, but these are not considered here.

The state space model (2.3)-(2.4) is a very general formulation. It is characterized by Markovian dynamics, conditionally independent observations, and mutually independent system and observation noise (Ristic et al. 2004). It allows for nonlinear models and non-Gaussian error processes. The usual goal is to estimate the state $\boldsymbol{x}_t$ and its density function over time. Sequential Monte Carlo methods, such as particle filters, provide standard solution techniques. However, the usual approaches do not provide estimates for the parameters, $\boldsymbol{\theta}_t$.

To estimate parameters in a state space model, two main approaches are available as discussed in Section (2.3). The approach we take, and which is the focus of this paper, is state augmentation (Kitagawa 1998). Here, the parameters of interest are appended to the state vector forming an augmented state vector, and standard sequential Monte Carlo methods are used to determine parameter estimates.

Let the augmented state vector be $\tilde{\boldsymbol{x}}_t = (\boldsymbol{x}',\ \boldsymbol{\theta}')'_t$, where $'$ denotes transpose, and t denotes time increment t. By making the parameters in the process model (2.3) follow a

random walk, the augmented state evolution equation becomes

$$\begin{pmatrix} \boldsymbol{x}_t \\ \boldsymbol{\theta}_t \end{pmatrix} = \begin{pmatrix} \boldsymbol{f}(\boldsymbol{x}_{t-1}, \boldsymbol{\theta}_t) \\ \boldsymbol{\theta}_{t-1} \end{pmatrix} + \begin{pmatrix} \boldsymbol{n}_t \\ \boldsymbol{\nu}_t \end{pmatrix} , \tag{3.2}$$

or

$$\tilde{\boldsymbol{x}}_t = \boldsymbol{g}(\tilde{\boldsymbol{x}}_{t-1}) + \tilde{\boldsymbol{n}}_t , \tag{3.3}$$

where the augmented system noise $\tilde{\boldsymbol{n}}_t$ now includes both the system noise, $\boldsymbol{n}_t$, as well as the disturbance term, $\boldsymbol{\nu}_t$, for the random walk of $\boldsymbol{\theta}_t$. The operator $\boldsymbol{g}$ includes both the movement dynamics f and the random walk for its parameters. The augmented observation equation (2.2) is now

$$\boldsymbol{y}_t = (\boldsymbol{H} \;\; \mathbf{0}) \; \tilde{\boldsymbol{x}}_t \, ,$$

where $\mathbf{0}$ is a vector of zeros and implies that we cannot observe the parameters, $\boldsymbol{\theta}_t$, directly. The augmented state space model can then be solved using the same sequential Monte Carlo methods as for regular state space models, since this is nothing more than a reformulation of the state. Hence, dynamic, or continuously time-varying, parameters are straightforward to estimate using state augmentation. Static, or slowly time-varying, parameters are, however, more complex.

The central goal in this study is to estimate the time-varying movement parameters that have a direct link to fur seal behaviour. Hence, identification of the appropriate time scale for this estimation is a key feature. Standard application of the augmented state space model (2.23) with a parameter random walk having a constant variance, σ_ν^2 (c.f. Kitagawa 1998), means that the estimated time-varying movement parameters simply follow short period fluctuations in the data, which is not always desirable for analysis and inference. For example, for vertical dive data collected at two second resolution, this means that movement parameter variations can oscillate within individual dives. For meaningful behavioural inference, we seek to identify behaviour on a longer time scale, so that we characterize variations associated with collections or ensembles of dives. Hence, we re-formulate the parameter estimation problem to focus on this longer time scale that is more directly linked to behaviour, as outlined below.

The analysis here is therefore sequentially carried out on short, overlapping windows of the movement time series. Within each data window, fixed values for the parameters of the process model are estimated. By considering the time sequence of these static parameter

estimates over the duration of the data record, the time variation of the movement parameters on the appropriate time scale for behavioural inference can be estimated. This idea of evolving parameters is well-established for non-stationary statistics which vary over time (Priestley 2004, Ch. 11), and has the advantage that the investigator can control the time scale for behavioural parameter estimation by choosing an appropriate data window width.

The estimation procedures for parameters that are static within an analysis time window also relies on the state augmentation procedures outlined above. The main idea is that by successively reducing the variance of the disturbance term, σ_ν^2, in the parameter random walk, the movement parameter vector, $\boldsymbol{\theta}_t$, ends up being fixed at a particular value. One way this can be accomplished is by using a single pass of a particle filter over an analysis time window and reducing σ_ν^2 with increasing t, until σ_ν^2 reaches a small value at the end of the analysis window and the parameters, $\boldsymbol{\theta}$, are locked in at their estimated values (Kitagawa 1998, Liu and West 2001). As a related alternative, Ionides et al. (2006) suggest using multiple passes of a state augmented particle filter, where each pass has a successively smaller σ_ν^2. They show this can yield the maximum likelihood value for $\boldsymbol{\theta}$. We make use of a hybrid approach, using the multiple iterated filtering approach of Ionides et al. (2006) algorithm, but using unweighted averages at each pass of the particle filter and a simple ramp-down of the variance, σ_ν^2. This is done within a time windowing sequential procedure and permits slowly time varying parameters to be estimated over successive windows. The goal is to interpret these movement parameters as they slowly vary in time at a scale that is sensitive to changes in northern fur seal behaviour.

3.1.3 Data

Here, we focus on analysing the dive record from a single female northern fur seal's track recorded on August 18, 2006. Figure (3.1) shows the depth measurements and the derived vertical velocity (via differencing the tag's depth channel) from the dive record for this single day selected from her 11.2 day foraging bout. The two-second sampling interval of the archival tag yielded 43,200 data points for this single day (1800 points per hour), and these data are the basis for our application. This data channel was selected for the movement analysis as it minimized some of the more complex error structures associated with the tag's speed channel.

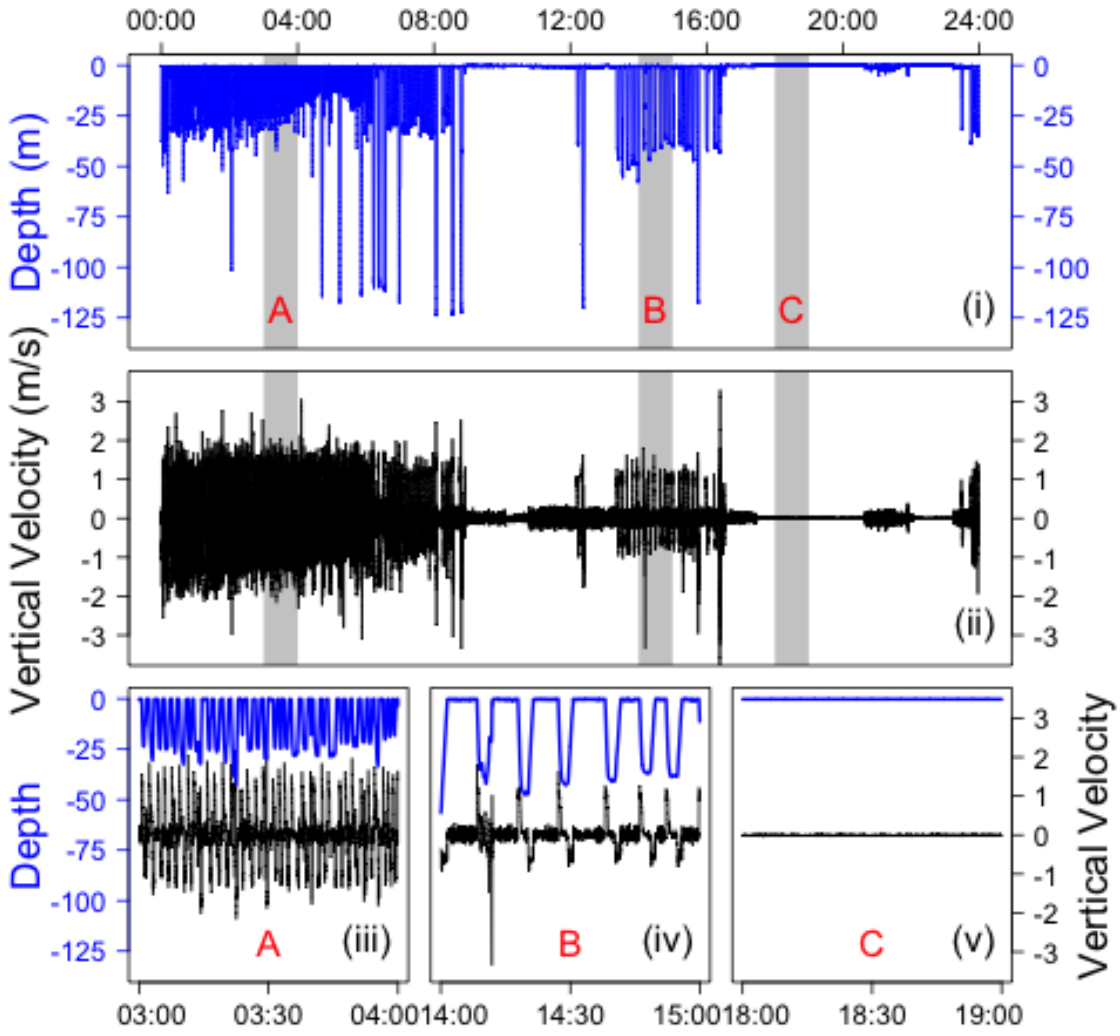

Figure 3.1: Observed time series of depth (panel (i), blue line), and derived vertical velocity (panel (ii), red line) from a northern fur seal collected on August 18^{th}, 2006, with hour of day indicated. There are three selected one hour long data segments, shaded in gray and labeled in green as A, B, and C. These periods will be referenced repeatedly with respect to the movement analysis and northern fur seal behaviour.

These data clearly show a distinction between dive and non-dive (surface swimming or resting) periods. The active dive periods also have distinctly different signatures or

behaviours. Figure (3.2) shows details of three selected one hour segments. In Segment A, we see rapid, regular, relatively shallow diving with large frequent vertical speed changes, which might be indicative of foraging/feeding. In Segment B, the dives are becoming deeper and less frequent and regular with more surface time and smaller vertical speeds, perhaps indicating exploring/searching. In Segment C, the fur seal is at the surface with a small vertical speed, and is either transiting or resting. The statistical character of these velocity observations obviously changes over the course of the record, and our goal is to quantify these by determining the associated (slowly) time-varying behavioural parameters of a suitable movement model.

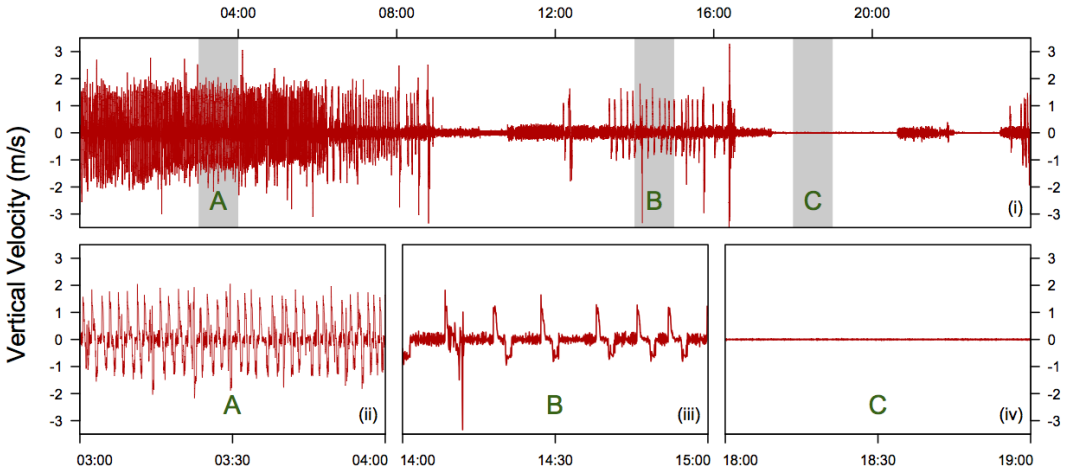

Figure 3.2: The original 24 hour time series and three details of the observed vertical velocity data [Figure 3.1.(ii)]. The three lower panels (panels (ii) - (iv)) highlight three 1 hour long data segments (1800 2-second time points) which correspond to the three sections of the 24 hour time series in panel (i) labeled as A, B, C.

A key aspect of state space modeling is to identify a suitable state evolution equation (2.1) or, here, the movement model, including a functional form for the dynamics and the statistics of the random forcing terms, $\boldsymbol{\nu}_t$.

A non-stationary time series analysis was undertaken in the form of the sample evolutionary auto-covariance function (ACVF) of the vertical velocity data. This is a way to account for auto-correlation that varies over the data record, analogous to time-freqency

analysis (e.g. Wittemyer et al. 2008, Polanksy et al. 2010). This analysis tells us whether or not a substantial linear relation exists between the series and its own lagged values, and what lag leads to the best predictability. Specifically, the ACVF was computed for 110 sliding time windows with a length of 26 minutes, overlapping the previous window by 13 minutes and covering the length of our selected 24-hour period. The idea is that within a single time window, the time series is approximately stationary and by considering the time sequence of ACVF, we can see how these statistics evolve over the analysis period. The time window of 26 minutes (or 780 data points) was chosen as a balance where the statistics of fur seal movement would be approximately stationary, but also that the data series covered enough individual dives so that the emergent behaviour of the fur seal could be determined.

Figure 3.3 (i) shows the evolutionary ACVF for observed vertical velocity, $\boldsymbol{y}$. It shows an obvious cycling between dive periods and quiescent non-dive periods. Within the dive period (e.g. between 00:00-06:00 hours, including Segment A), we see an oscillating and decaying ACVF with a relatively large variance, suggesting a periodic process. During the non-dive periods (e.g. 17:00-22:00, including Segment C), the ACVF has low variance and decays to zero for lags one and beyond, suggesting a purely white noise process. Behavioral distinctions between the dive and non-dive periods are evident, especially in terms of the shifting in the magnitude of the velocity variance. The periods of diving show a dominance of temporally coherent motions with periods reflecting the dive length.

Because direct observation of northern fur seal at-sea behaviour is not possible, we must rely on some kind of movement model to detect, quantify and categorize the time series tag data into meaningful patterns for inference. These movement models can take on a wide variety of forms (e.g. Codling et al. 2008) and can exhibit a range of non-stationary behaviour; our goal was to use a movement model to infer behaviour such as active foraging in a window of northern fur seal tag data.

Identification of a suitable movement model is undertaken next. We considered the candidate class of models to be the auto-regressive moving-average (ARMA) models based on the qualitative nature of the ACVF analysis. This class of models can be equated to the dynamic systems framework as shown in the following section.

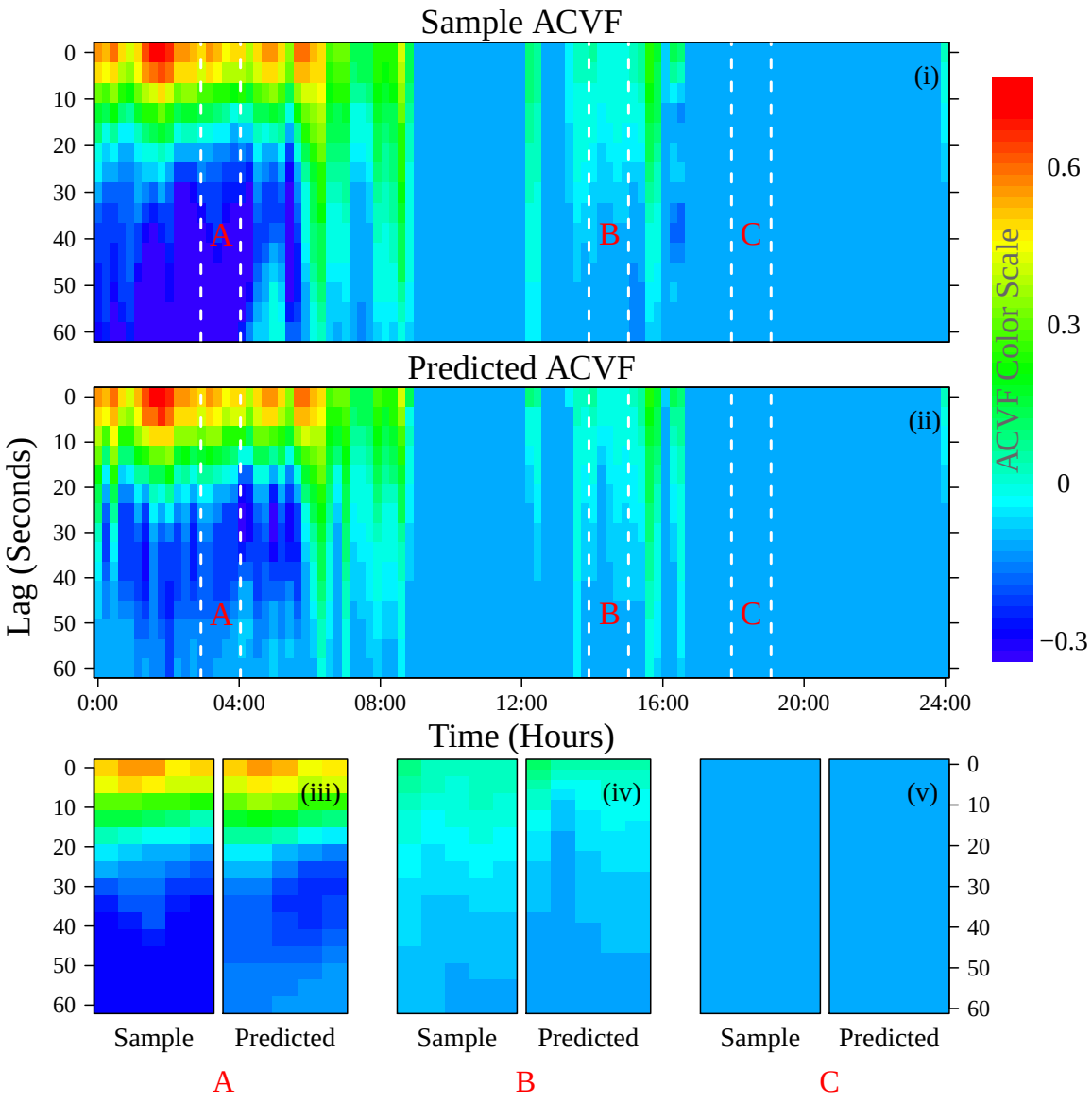

Figure 3.3: Evolutionary auto-covariance function (ACVF) of vertical velocity. Panel (i) shows the sample ACVF computed from the data with time of day indicated. Panel (ii) shows the corresponding ACVF predicted from the estimated behavioural parameters of the movement model. Panels (iii)-(v) show the details for time segments A, B, and C (indicated by dotted lines in panels (i) and (ii)) for both the sample and predicted ACVF.

3.1.4 AR(p) Process and the Movement Model

Suppose $\boldsymbol{X}_t$ is a sequence of states up to time t, $\{\boldsymbol{x}_0, \boldsymbol{x}_1, \ldots, \boldsymbol{x}_t\}$. A zero-mean AR($p$) process for the state at time t satisfies the difference equation

$$x_t = a_1 x_{t-1} + a_2 x_{t-2} + \ldots + a_p x_{t-p} + \eta_t \tag{3.4}$$

where a_p denotes the p^{th}-order autoregression parameter. Here η_t (with mean 0) denotes the (assumed additive) system noise or error, and controls the amplitude of the forcing (i.e. the degree to which the signal is obscured depends on the amplitude of the signal and the magnitude of η_t).

Using formal model fitting procedures based on Bayesian and Akaike information criteria, an auto-regressive model of order two (ARMA(2,0) or AR(2)) was chosen as the most suitable model,

$$x_t = a_1 x_{t-1} + a_2 x_{t-2} + \eta_t. \tag{3.5}$$

Here, x_t represents the vertical velocity at time t and becomes a component of the augmented state $\tilde{x}_t$. What follows is the derivation of the AR(2) movement model of (2.5). Here the system model is a random walk model that involves two-stage dependence such that x_t, the state estimate of vertical speed at time t, depends linearly on the state in the previous two lagged time points, x_{t-1}, x_{t-2}, through a set of movement parameters (a_1, a_2). This AR(2) movement model is derived by discretizing a 2^{nd}-order linear ordinary differential equation (ODE) as follows :

$$\frac{d^2x}{dt^2} = \alpha_1 \frac{dx}{dt} + \alpha_2 x + \eta, \tag{3.6}$$

$$\begin{aligned}
\text{where} \quad \frac{dx}{dt} &= \lim_{\Delta t \to 0} \frac{1}{\Delta t} \left[x_t - x_{t-\Delta t} \right] \\
\text{and} \quad \frac{d^2x}{dt^2} &= \lim_{\Delta t \to 0} \left[\frac{1}{\Delta t} \left(\frac{1}{\Delta t} \left[x_t - x_{t-\Delta t} \right] - \frac{1}{\Delta t} \left[x_{t-\Delta t} - x_{t-2\Delta t} \right] \right) \right] \\
&= \lim_{\Delta t \to 0} \frac{1}{(\Delta t)^2} \left(x_t - 2x_{t-\Delta t} + x_{t-2\Delta t} \right) \ .
\end{aligned} \tag{3.7}$$

By substituting these discretized approximations of $\frac{dx}{dt}$ and $\frac{d^2x}{dt^2}$ into (3.6), we can solve for the vertical velocity at the current time, x_t, as a function of the velocity at the two previous

time increments. Substituting $\Delta t = 1$, representing one time increment, gives $t - \Delta t = t - 1$, and $t - 2\Delta t = t - 2$, and (3.6) is rewritten as

$$x_t - 2x_{t-1} + x_{t-2} \quad = \quad \alpha_1 \left([x_t - x_{t-1}]\right) + \alpha_2 x_t + \eta_t. \tag{3.8}$$

Therefore, we can rearrange (3.8) into an AR(2) process (assuming $\alpha_1 + \alpha_2 \neq 1$)

$$x_t \quad = \quad \frac{(2-\alpha_1)}{1-\alpha_1-\alpha_2}\, x_{t-1} - \frac{1}{1-\alpha_1-\alpha_2}\, x_{t-2} + \frac{1}{1-\alpha_1-\alpha_2}\, \eta_t.$$

By letting

$$a_1 = \left(\frac{2-\alpha_1}{1-\alpha_1-\alpha_2}\right), \quad \text{and } a_2 = \left(\frac{-1}{1-\alpha_1-\alpha_2}\right),$$

gives an AR(2) process

$$x_t \quad = \quad a_1 x_{t-1} + a_2 x_{t-2} + \tilde{\eta}_t \tag{3.9}$$

where a_1, a_2 are parameters and $\tilde{\eta}_t$ is a purely random Gaussian process of independent identically distributed random variables. We assume without loss of generality that $E(\tilde{\eta}_t) = 0$ for all t, and $Var(\tilde{\eta}_t) = \sigma_\eta^2$. Figure (2.1) of the previous Chapter shows a realization of 1800 observations (or 1 hour of tag data) from an AR(2) model, $x_t = 1.5x_{t-1} - 0.6x_{t-2} + \tilde{\eta}_t$, with $\tilde{\eta}_t \sim \mathcal{N}(0, 0.21)$.

Thus, we have a simple mathematical representation of a discrete time-dependent system for northern fur seal vertical speed. The AR(2) model of (3.9) is a difference equation, which corresponds in continuous time to a second-order, stochastic, ordinary differential equation. It is different from linear regression in that the states x_{t-1}, x_{t-2} are random variables, and x appears on both sides of the equation.

The real advantage of selecting Equation (3.9) as the state evolution model, or movement process model, is the parameters a_1, and a_2 are interpretable in terms of fur seal dive behaviour. If, say, $a_1 = a_2 = 0$ then the model corresponds to a purely white noise process which might describe incoherent vertical velocity signal associated with surface swimming. If $a_1 = 1$, and $a_2 = 0$, it is a simple random walk. Generally, however, both a_1 and a_2 will be non-zero. If $a_2 < -a_1^2/4$ movement is pseudo-periodic and could describe a regular and repeating set of dives (Priestley 2004; section 3.5.3). The vertical motion is aperiodic when $a_2 > -a_1^2/4$. The variance of $\tilde{\boldsymbol{\eta}}_t$ scales the magnitude of the dive velocity. The model in Equation (3.9) is therefore a flexible description for fur seal vertical motion and its parameter values can be linked to behaviour.

3.1.5 Application

The state augmentation approach estimates the parameters a_1 and a_2 of the movement model (3.5). However the model must first be in the Markovian form (2.3) or (2.23), i.e. with a dependence on only one time lag. We therefore rewrite (3.5) as follows:

$$\begin{pmatrix} x_t \\ \zeta_t \end{pmatrix} = \begin{pmatrix} a_1 & a_2 \\ 1 & 0 \end{pmatrix} \begin{pmatrix} x_{t-1} \\ \zeta_{t-1} \end{pmatrix} + \begin{pmatrix} \epsilon_t \\ 0 \end{pmatrix} \tag{3.10}$$

where the dummy variable ζ has been introduced to transform a non-Markovian model into a Markovian one; i.e., $\zeta_t = x_{t-1}$.

The original state (the left-hand-side of (3.10)) is next augmented with the parameters a_1 and a_2. As in (2.23), these follow a random walk

$$a_{1,t} = a_{1,t-1} + \nu_{1,t},\ \nu_{1,t}\ \sim \mathcal{N}(0, \sigma_\nu^2), \qquad a_{2,t} = a_{2,t-1} + \nu_{2,t},\ \nu_{2,t}\ \sim \mathcal{N}(0, \sigma_\nu^2) \tag{3.11}$$

where the variance of the disturbance term, σ_ν^2, must be specified appropriately to allow for estimation of the movement parameters (See off-line estimation procedure in Appendix A).

The augmented state vector is therefore $\tilde{\boldsymbol{x}}_t = (x_t,\ \zeta_t,\ a_{1,t},\ a_{2,t})'$, and the state evolution equation is defined by (3.10) and (3.11). The augmented system noise term is $\tilde{\boldsymbol{n}}_t = (\epsilon_t,\ 0,\ \nu_{1,t},\ \nu_{2,t})'$. Because of the estimation of the movement parameters, a_1 and a_2, inside the augmented state vector, this is now a non-linear state evolution equation.

The vertical velocity observations [Figure 3.2 (i)] are contained in $\boldsymbol{y}_t$ (Equation 3.1), and it is assumed that the observation error, e_t, is an independently distributed $\mathcal{N}(0, \sigma_{o,t}^2)$ random variable. The augmented observation operator in (2.2) is a row vector $\tilde{\boldsymbol{H}} = (1\ 0\ 0\ 0)$ which multiplies the augmented state $\tilde{\boldsymbol{x}}_t$, and indicates we observe only its first element.

The statistics of the observation error, e_t, and the system noise, ϵ_t, are difficult to estimate since stochastic variations seen in the observations are due to both measurement noise and fluctuations due to the animal movement process itself. Their variances can, however, be separated using the ACVF and the principle that the observation error is uncorrelated over time, whereas animal movement is time-correlated (see Appendix A). The advantage of estimating the system and observation error variances off-line is that it minimizes the number of identifiable parameters that need to be estimated in the state augmentation procedure. Note too, that the system noise variance, σ_ϵ^2, is a key element in the movement model, and acts by scaling its overall magnitude.

As part of the implementation of the state-space model, the system noise takes the form of a normal mixture process,

$$\epsilon_t \sim c_1 N(0, \sigma^2_{\epsilon,t}) + c_2 N(0, \delta\sigma^2_{\epsilon,t}). \tag{3.12}$$

The system noise has an overall variance corresponding to the off-line estimated σ^2_ϵ. We chose $c_1 = 0.9$, $c_2 = 0.1$, and $\delta = 10$ which allow the system noise variance to have occasional large values. This mixture-error form for the system error allows the state estimates to closely follow the observations when abrupt changes in velocity occur, such as those found at the onset of a dive.

The state x_t (and $\zeta_t = x_{t-1}$), as well as the parameters a_1 and a_2 are jointly estimated using a state-augmented particle filter based on sequential importance resampling (Section 2.3). This estimation is carried out for each of the 110 analysis time windows (each 26 minutes long, and overlapping by 13 minutes). The slow time variation in these parameters across the 110 windows corresponds to the changes in fur seal behaviour over the data record.

To estimate the static parameters a_1 and a_2 within a single time window, we use the following algorithm adapted from Ionides et al. (2006):

1. Run the particle filter with the data y_t using the state augmented model (3.10)-(3.11), and with system noise and observation noise as specified above. Set the initial random walk variance for the parameters at $\sigma^2_\nu(0)$ (the bracketed value indicates the iteration number), as well as initial values for $a_{1,t=0}(0)$ and $a_{2,t=0}(0)$. Take the mean estimated by the particle filter across all time increments of $\hat{a}_{1,t}(0)$ and $\hat{a}_{2,t}(0)$ to be the parameter values for iteration zero, i.e. $\hat{a}_1(0)$ and $\hat{a}_2(0)$.

2. For iterations $k = 1, \ldots, m$. Let $\sigma^2_\nu(k) = \alpha\sigma^2_\nu(k-1)$, where α is a discounting factor used to control the reduction in the random walk variance at each iteration. Use as initial values for $a_{1,t=0}(k) = \hat{a}_1(k-1)$ and $a_{2,t=0}(k) = \hat{a}_2(k-1)$. Take the mean of $a_{1,t}(k)$ and $a_{2,t}(k)$ from the particle filter for estimates $\hat{a}_1(k)$ and $\hat{a}_2(k)$.

3. Stop after m iterations. Use $\hat{a}_1(m)$ and $\hat{a}_2(m)$ as final estimates for a_1 and a_2.

This is a simplified (unweighted) version of the multiple iterated filter of Ionides et al. (2006). It was found that the following values yielded good results: (i) an initial $\sigma^2_\nu(0)$ = $(0.1 \text{ m/s})^2$; (ii) a discounting factor of $\alpha = 0.5$, and (iii) $m = 10$ iterations provided a

stopping criterion as $\sigma_\nu^2(10)$ was by then small enough that the parameters were effectively fixed at a constant value. This above algorithm was applied for each of the 110 time-analysis windows to estimate the time variation of a_1 and a_2.

To test the accuracy and stability of this algorithm, we ran multiple runs with synthetic data generated using the movement model and known parameters. These tests indicated that the above algorithm provided a good recovery of the movement parameters.

3.1.6 Results

State estimates $\boldsymbol{x}$ for the vertical velocity are obtained as part of the analysis but are not of interest in our behavioural inference. The estimated state (Figure 3.4 (ii), and (iii), in grey) conforms well to the velocity observations (3.1(ii),3.2(i)). In fact, it is simply a low pass filtered version, which follows the underlying velocity signal. The state estimates are not discussed further.

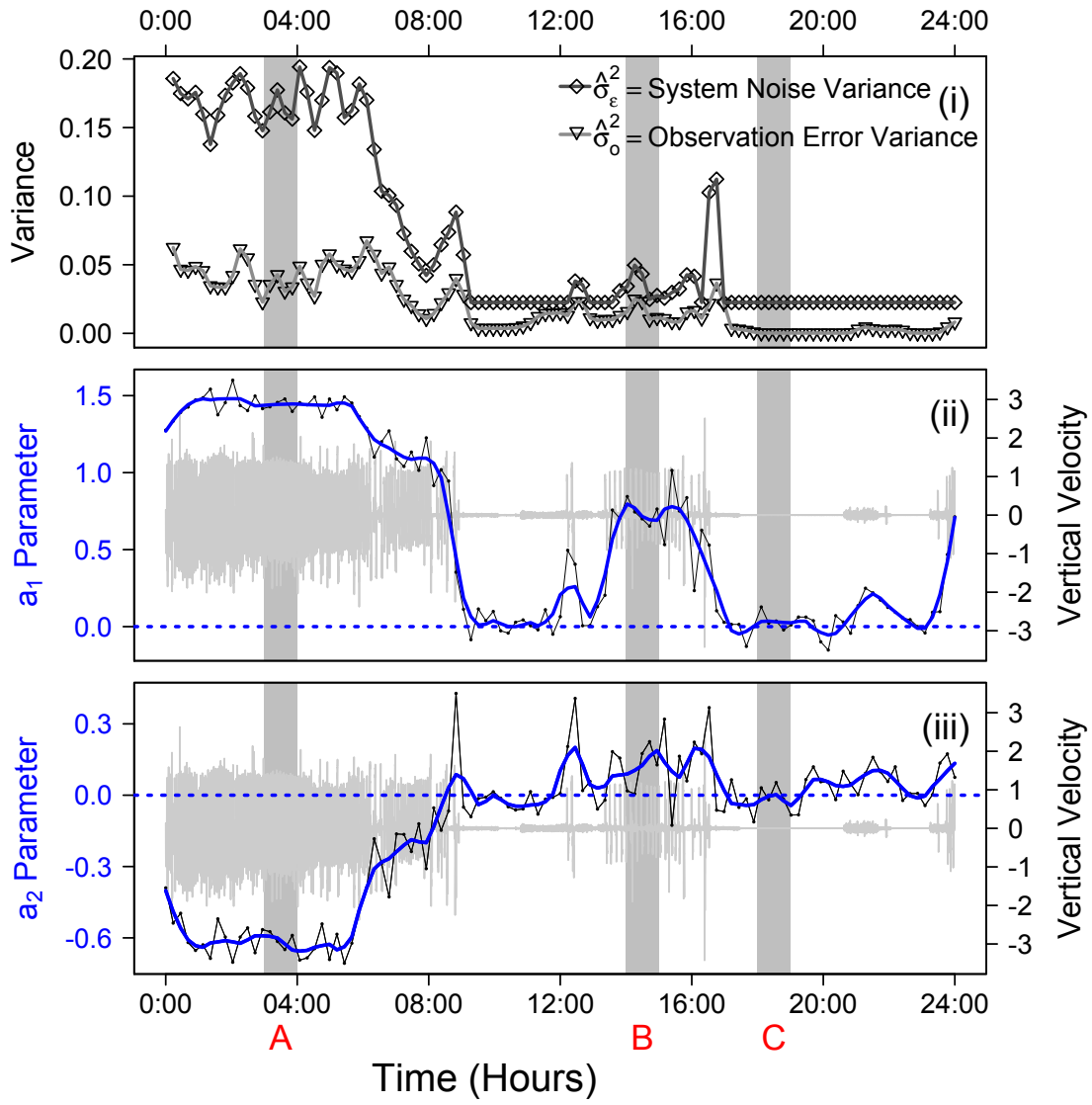

Figure 3.4: Parameter estimates over time as obtained from the analysis. Panel (i) shows the estimated observation error variance and system noise variance. Panel (ii) shows the estimated a_1 coefficients from the state-space model (black line), along with a low-pass filtered version of it (blue line). Panel (iii) shows the corresponding information for the a_2 coefficient. The state estimate for the vertical velocity time series is also shown (light grey). The detailed data segments (A, B, C) corresponding to equivalent segments in Figures (3.1, 3.2, 3.3) are indicated by grey shading.

Figure 3.4 (i) shows the off-line estimates of the system noise variance and observation error variance, and how they evolve through time for each of the analysis time windows. Both scale proportionally to vertical speed, and large during periods of active diving, and drop towards zero during non-diving phases. These variances were used as inputs, σ_o^2 and σ_ϵ^2, to the state space model.

Figure 3.4 also shows the estimated slowly time varying movement parameters from the analysis, i.e. the auto-regressive coefficients a_1 [Figure 3.3 (ii)] and a_2 [Figure 3.3 (iii)]. During the initial part of the day [00:00 - 06:00, and highlighted in Segment A, Figure 3.2 (ii), and Segment A, Figure 3.4 (i)-(iii)], the fur seal exhibits rapid, frequent, regular dives. For this time interval, $a_1 \approx 1.5$, $a_2 \approx -0.6$, and the system noise variance, σ_ϵ^2, is high (0.15 - 0.2). Just past mid-day, there is a 4 hour period of occasional deep dives with large velocities [highlighted in Segment B, Figure 3.2 (iii), and Segment B, Figure 3.4 (i)-(iii)]. The system noise variance increases to 0.05, $a_1 \approx 0.8$ and $a_2 \approx 0.1$. There are also periods where the fur seal remains at the surface with small vertical speeds [highlighted in Segment C, Figure 3.2 (iv), and Segment C, Figure 3.4 (i)-(iii)] in which the a_1 and a_2 parameters and system noise are near zero. The behavioural signatures in these vertical velocity data are however not always in discrete modes, but have time variations and mixes of behaviours. These are captured well by the continuously time-varying movement parameters [Figure 3.4 (ii), (iii)].

To further validate the results, we address the question of how well the estimated parameters can re-construct the emergent statistical properties of the data. The evolutionary ACVF can be predicted from knowledge of the time-evolution of a_1, a_2 and σ_ϵ^2 (Priestley 2004, Ch. 3-4). The predicted ACVF can then be compared to its corresponding sample version computed from the data in Figure 3.3 (ii). Note that the sample ACVF does not separate out stochastic fluctuations in the movement process from the observations, as does the state space model. Also, since the movement model is a simplified representation of reality it can only capture a portion of the variability in the data. Hence we do not expect exact comparability or reproduction. It is nevertheless a useful comparison and validation.

The main features of the predicted ACVF [Figure 3.3 (ii)] compare very well with the sample ACVF [Figure 3.3 (i)], especially at the important small lags. In particular, the oscillation and decay of the ACVF with increasing lag is predicted for the initial foraging period [Segment A, Figure 3.3 (iii)]. In fact, the major differences for this period are the added noise at large lags in the sample ACVF, and that the sample ACVF has deeper

negative values at larger lags than does the predicted one. This latter discrepancy occurs since the observed velocity series does not have an exact periodicity even within a single time analysis window. That is, the dives are regular, but not exactly repeating. In Figure 3.3 (iv) , the sample and predicted ACVF are compared for Segment B, characterized by irregular diving; the magnitude and decay rate of the ACVF is well captured in both sample and predicted ACVF. Finally, in [Figure 3.3 (v)] , the low velocities and cutoff after lag zero characterizing Segment C are seen in both the sample and predicted values.

3.1.7 Concluding Remarks

We have explored the idea of estimating behaviour parameters from marine animal archival tag data together with movement models. The central idea is that by estimating the time variation of parameters for a suitable movement model, researchers can then objectively and quantitatively infer animal activity (and its behavioural state). Here, we have offered a statistical-dynamical approach suitable for extracting behavioural information from high resolution data, and one that is widely applicable to a variety of movement models and observation types. It offers an alternative to behavioural switching state space models (Jonsen et al. 2007, Patterson et al. 2009), and is flexible enough to obtain solutions even for nonlinear and non-Gaussian cases. Our approach draws from the state augmentation procedures of Kitagawa (1998) and Ionides et al. (2006).

Our application focused on analysis and interpretation of the vertical movement data from a tagged northern fur seal. The analysis provides slowly time-varying estimates for the movement parameters. The application also highlighted many issues which are expected to arise in any analysis of high resolution movement data. They include: (i) the identification of an appropriate movement model; (ii) choosing the appropriate time scale for parameter estimation; and (iii) the estimation of identifiable parameters. Remarks on each of these issues are offered below.

Identification of the appropriate movement model is an important issue that affects interpretation of the data, and is key to making meaningful conclusions about behaviour. In fact, we want models for which the parameters are directly interpretable as behaviour. Model choice must be based on theoretical precepts, as well as statistical features of the data. For this study, the discrete-time AR(2) model was well-suited for the vertical velocity data, but continuous time formulations of movement are also possible within the state-space framework (e.g. Johnson et al. 2008). Generalizing the approach to two-dimensional

and three-dimensional movement world require different types of models, such as correlated random walks (Morales et al. 2004). These would replace the process model, and parameters could be estimated with the state augmentation procedure.

An issue specific to the method is the choice of the time-scale for the movement analysis. The lower limit for time window length is chosen based on the natural time scale of the movement model. Here, the AR(2) model is applicable to a time series comprised of a coherent collection of similar dives, and not a single dive. Therefore, in the northern fur seal, the window width must be at least wide enough to encompass a series of dives. The upper limit on the window length is based on the time-scale at which fur seal behavioural changes take place, with the idea that the parameter estimation uses a data window that is approximately stationary (i.e., parameters should be constant over the window). The 26 minute window was a reasonable compromise, and the 13 minute overlap allows for us to resolve more abrupt changes in the parameter values. Note that some types of models with low resolution data for which behaviour can be directly inferred from the state may not need to be time windowed and state augmentation can proceed directly without recourse to the multiple iterated filtering.

Parameter identifiability is a concern for many ecological and movement models, especially as complexity is increased. One possibility is to specify well-defined values for the magnitude and form of the process and observation errors. Another is the estimation of the minimal number of identifiable parameters, and determining parameters off-line wherever possible. Alternatively, hybrid particle-MCMC approaches (e.g. Andrieu et al. 2010) could be considered.

In summary, recent review articles have identified state space models (Patterson et al. 2008) and hierarchical Bayesian approaches (Schick et al. 2008) as two important directions for extracting ecologically meaningful information from animal tag data. Here we have used a particle filter approach for an augmented state-space model that is designed for the type of large volume, high resolution motion time series recorded by archival animal tags. The approach of this study is a promising direction for fully exploiting the behavioural information in these rich data sets.

3.2 Literature Cited

Andrieu, C., A. Doucet, and R. Holenstein. 2010. Particle Markov chain Monte Carlo. Journal of the Royal Statistical Society B, 72(3): 269-342.

Bodkin, J.L., D.H. Monson, and G.G. Esslinger. 2007. Activity Budgets Derived From Time-Depth Recorders in a Diving Mammal. Journal of Wildlife Management, 71(6): 2034-2044.

Codling, E.A., M.J. Plank, and S. Benhamou. 2008. Random walks in biology. Journal of the Royal Society Interface, 5: 813-834.

Ionides, E.L., C. Breto, and A.A. King. 2006. Inference for nonlinear dynamical systems. Proceedings of the National Academy of Sciences, 103: 18438-18443.

Johnson, D.S., J.M. London, M.-A. Lea, and J.W. Durban. 2008. Continuous-time correlated random walks model for animal telemetry data. Ecology, 89: 1208-1215.

Jonsen, I.D., R.A. Myers, and J.M. Flemming. 2005. Robust state-space modeling of animal movement data. Ecology, 86: 2874-2880.

Jonsen, I.D., R.A. Myers, and M.C. James. 2007. Identifying leatherback turtle foraging behaviour from satellite-telemetry using a switching state-space model. Marine Ecology Progress Series, 337: 255-264.

Liu, J., and M. West. 2001. Combined parameter and state estimation in simulation based filtering. In Sequential Monte Carlo Methods in Practice, Doucet, A., N. de Freitas, and N. Gordon (eds); Springer: New York; pp. 197-217.

Kitagawa, G. 1998. A self-organising state-space model. Journal of the American Statistical Association, 93: 1203-1215.

Morales, J.M., D.T. Haydon, J. Friar, K.E. Hosinger, and J.M. Fryxell. 2004. Extracting more from relocation data: building movement models as mixtures of random walks. Ecology, 85(9): 2436-2445.

Patterson, T.A., L. Thomas, C. Wilcox, O. Ovaskainen, and J. Matthiopoulos. 2008. State-space models of individual animal movement. Trends in Ecology and Evolution, 23: 87-94.

Patterson, T.A., M. Basson, M.V. Bravington, and J.S. Gunn. 2009. Classifying movement behaviour in relation to environmental conditions using hidden Markov models. Journal of Animal Ecology, 78: 1113-1123.

Polansky, L., G. Wittemyer, P.C. Cross, C.J. Tambling, and W.M. Getz. 2010. From moonlight to movement and synchronized randomness: Fourier and wavelet analyses of animal location time series data. Ecology, 91: 1506-1518.

Priestley, M.B. 2004. Spectral Analysis and Time Series. Academic Press. London. 890pp.

Ristic, B., S. Arulampalam, and N. Gordon. 2004. Beyond the Kalman Filter: Particle Filters for Tracking Applications. Artech House: Boston, 318pp.

Ropert-Coudert, Y., and R.P. Wilson. 2005. Trends and perspectives in animal-attached remote sensing. Frontiers in Ecology and the Environment, 3(8): 437-444.

Schick, R.S., S.R. Loarie, F. Colchero, B.D. Best, A. Boustany, D.A. Conde, P.N. Halpin, L.N. Joppa, C.M. McClellan, and J.S. Clark. 2008. Understanding movement data and movement processes: current and emerging directions. Ecology Letters, 11: 1338-1350.

Wittemyer, G., L. Polansky, I. Douglas-Hamilton, and W.M. Getz. 2008. Disentangling the effects of forage, social rank, and risk on movement autocorrelation of elephants using Fourier and wavelet analyses. Proceedings of the National Academy of Sciences USA, 105: 19108-19113.

Chapter 4

Hierarchical Bayes methods for modelling uncertainty in at-sea animal behaviour

Scaling up movement behaviour to the landscape level is of key importance in understanding the processes that limit northern fur seal success in the Bering Sea. The main barrier to developing models addressing these landscape level questions is the vast difference in scale between the at-sea behaviour of the individual seal and the oceanic environment through which that fur seal passes. The environment is typically only observed incompletely and with large and unknown amounts of measurement error and data uncertainty. As well, the environment can only be approximated by combining various data sources, often at very different scales.

We introduce in this chapter the hierarchical Bayesian model as a means to account for multiple sources of uncertainty in the measurement and understanding of the fur seal's at-sea environment. We do not set out to identify which sections of the Bering Sea might be of conservation priority with respect to foraging hotspots, and therefore have not approached this as a spatial problem. Instead our model is constructed with the idea that a fur seal swims continuously through a variety of habitats, in some of which the fur seal engages in repetitive dive activity, in other sections where it engages in consistent but infrequent dive behaviour, and still other sections in which it doesn't dive at all. Our goal is to understand if there was something in the variable field through which the fur seal swam, that influenced

the observed fur seal's behaviour.

This chapter is motivated by the problem of (i) processing a dataset collected from satellite-linked, archival tagged female northern fur seals with pups, and (ii) understanding their at-sea behaviour in response to changes in the environment. The archival tag dataset that underlies the approach we present in this chapter was collected from a set of eleven female fur seals nourishing pups at Reef Rookery, on St. Paul Island in the Pribilof Islands, Alaska. Each fur seal was tagged with an archival tag that measured a series of in-situ variables at two-second intervals, as well as a lower-resolution satellite GPS tag from which the fur seal's location could be determined. The high-resolution depth channel was differenced from each seal's archival tag to create a measure of "vertical velocity" that serves as the basis for the application of an augmented state space model to describe fur seal movement. These state space methods are ideal for the application where fur seal movement and behaviour evolves slowly over time, and were presented in Chapter (2) and discussed in detail in Chapter (3).

In this chapter, we begin by reviewing Bayesian hierarchical regression as a framework for modelling northern fur seal behaviour as output from an augmented state space model. We then introduce methods to account for uncertainty in the measurement of their at-sea environment, and show how to incorporate this into the hierarchical Bayesian framework. We explore the strengths and weaknesses of the hierarchical Bayesian model as a tool for inference about at-sea seal behaviour, and end with a discussion of other approaches that could be used in this context.

4.1 Model Development

4.1.1 Bayesian Regression

There are two important components for carrying out statistical inference in a Bayesian regression setting; the likelihood function of the data

$$P(Data|Parameters)$$

and the prior densities for the model unknowns

$$P(Parameters).$$

Consider a data set collected from a single individual. The information provided by the data is incorporated by selecting a statistical model that reflects our beliefs about the data which takes the form of a likelihood function, e.g., $P(\boldsymbol{y}|\boldsymbol{\theta})$, where $\boldsymbol{y}$ denotes the dependent variable, and $\boldsymbol{\theta}$, any parameters associated with the regression of $\boldsymbol{y}$. Bayesian models have an inherently hierarchical structure. Parameters are treated as if they are random variables and are given *prior* distributions described by a set of prior parameters, $\boldsymbol{\psi}$, and can be written to reflect this hierarchy, i.e. $P(\boldsymbol{\theta}|\boldsymbol{\psi})$. This formulation emphasizes that the prior distribution of $\boldsymbol{\theta}$ is itself characterized by a set of distribution parameters, $\boldsymbol{\psi}$. The choice of prior distribution even among those that are intended to be non-informative should be done carefully, particularly in high-dimensional problems (Lambert et al. 2005).

After observing the data, we can conceptually update our prior beliefs using Bayes' Theorem to obtain the *posterior* distribution of the regression parameters, i.e. $P(\boldsymbol{\theta}|\boldsymbol{y})$ or

$$P(Parameters|Data).$$

Bayesian inference is based upon this posterior density and is done by combining the likelihood of the observed data, and prior densities as follows

$$P(\boldsymbol{\theta}|\boldsymbol{y}) = \frac{P(\boldsymbol{\theta}, \boldsymbol{y})}{P(\boldsymbol{y})} = \frac{P(\boldsymbol{y}|\boldsymbol{\theta})P(\boldsymbol{\theta})}{\int P(\boldsymbol{y}|\boldsymbol{\theta})P(\boldsymbol{\theta})\ d\boldsymbol{\theta}}. \tag{4.1}$$

Once the data $\boldsymbol{y}$ are observed, the denominator of (4.1) is constant with respect to $\boldsymbol{\theta}$, and represents the marginal distribution of the data, i.e.

$$\int P(\boldsymbol{y}|\boldsymbol{\theta})P(\boldsymbol{\theta})\ d\boldsymbol{\theta} = \int P(\boldsymbol{y}, \boldsymbol{\theta})\ d\boldsymbol{\theta} = P(\boldsymbol{y}).$$

Therefore the shape of the posterior is determined entirely by the numerator of (4.1), and we emphasize this by rewriting Equation (4.1) as

$$P(\boldsymbol{\theta}|\boldsymbol{y}) \propto P(\boldsymbol{y}|\boldsymbol{\theta})P(\boldsymbol{\theta}). \tag{4.2}$$

Bayesian methods then proceed by using sample-based methods for approximating unknowns in the relation (4.2).

4.1.2 Bayesian hierarchical model

The Bayesian hierarchical model is a statistical methodology for handling complex problems by synthesizing information about a system in a coherent manner (Cressie et al. 2009).

Within the Bayesian framework, we rely on probability theory to provide decompositions of complex joint probability distributions into a series of simpler conditional statistical models that aim to explain the relationship between data and a causal process. The strength of this approach is that it can use decompositions that reflect the causal mechanisms, and our uncertainty around that causal process. In this chapter we present a flexible 3-stage hierarchical model in which the general form of the hierarchical model is described as follows (adapted from Berliner 1996, and Cressie et al. 2009).
The first level of the hierarchy is the data model or likelihood function that describes the observation process

$$P(Data|Process, Parameters), \tag{4.3}$$

the second level is a process model that describes the state process

$$P(Process|Parameters), \tag{4.4}$$

and the third level is a parameter model that expresses our uncertainty in the parameters describing the process model

$$P(Parameters). \tag{4.5}$$

Collectively these three conditional models (Equations 4.3, 4.4, 4.5) provide a means for updating the target distribution of the unknown $Parameters$ updated by the observed $Data$, i.e.,

$$P(Parameters|Data) \propto P(Data|Process, Parameters)P(Process|Parameters)P(Parameters) \tag{4.6}$$

The general approach to hierarchical modelling can be considered as a large set of stochastic formulations that include many popular models such as random effects models and generalized linear mixed models, and are flexible enough to solve a variety of inference problems (e.g. Gelman et al. 2004, Chapters 5, 15, and references cited therein). Here, we present how this framework can be adapted to the analysis of seal behaviour during an at-sea foraging voyage. Specifically, the analysis will focus on (i) the derivation of a set of qualitatively distinct seal behavioural classes based on model parameters and (ii) testing for relationships between these and the environmental covariates. We will formally acknowledge

the randomness in both seal behaviour and the environment process by building a hierarchy of conditional models to describe the complexity of our data and the process that generates animal behaviour (Cressie et al. 2009).

4.1.3 Identifying Behaviour Class

Recall from Section (3.1.4) of this thesis, that the AR(2) movement model was derived by discretizing a 2nd-order linear ordinary differential equation. The parameters, a_1, and a_2, of the AR(2) system model determine the qualitative behaviour of the process. We can derive this relationship through the roots of the quadratic characteristic equation of the difference equation.

We may obtain the conditions for asymptotic stationarity by studying the form of the general solution to the AR(2) difference equation; i.e.,

$$x_t \quad = \quad a_1 x_{t-1} + a_2 x_{t-2} + \tilde{\eta}_t. \tag{4.7}$$

Following Priestley (2004) with a slight change of notation, first consider a system without noise (i.e., with $\tilde{\eta}_t = 0$). Then, as with 2nd-order differential equations, the 2nd-order difference equation,

$$x_t + a_1 x_{t-1} + a_2 x_{t-2} = 0 \ , \tag{4.8}$$

will have a general solution

$$x_t = c_1 \mu_1^t + c_2 \mu_2^t \ , \tag{4.9}$$

where μ_1, and μ_2 are the roots (possibly complex or coincident) to the quadratic equation

$$f(\mu) = \mu^2 + a_1 \mu + a_2 \text{ (2004; Equation 3.5.23).} \tag{4.10}$$

From Priestley (2004), the AR(2) process is stationary if the general solution (Equation 4.9) decays to zero as $t \to \infty$, or equivalently when the two roots of $f(\mu) = 0$ lie inside the unit circle (Priestley 2004; Equations 3.5.26 and 3.5.28).

From Priestley (2004; Equation 3.5.29), the necessary and sufficient conditions for the roots to lie inside the unit circle are:

1. $a_1 + a_2 < 1$

2. $a_2 - a_1 < 1$

3. $|a_2| < 1$.

These conditions specify the interior of a triangular region of stationarity in the (a_1, a_2) parameter plane (Figure 4.1, Priestley 2004, Shumway and Stoffer 2006).

The roots of Equation (4.10) are given by

$$\mu_1, \mu_2 = \frac{1}{2}\left(a_1 \pm \sqrt{a_1^2 + 4a_2}\right).$$

If $a_1^2 + 4a_2 \geq 0$ in (4.11), then the roots μ_1, μ_2 are real. If, in addition, $a_1 > 0$ and $a_2 < 0$, then μ_1, μ_2 are positive and the the stationary process decays smoothly to zero; if $a_1 > 0$ and $a_2 > 0$, then μ_1, μ_2 are of opposite signs; and if $a_1 < 0$ and $a_2 < 0$, then both μ_1, μ_2 are negative, and the stationary process alternates in sign as it decays.

If $a_1^2 + 4a_2 < 0$ in (4.11), then $-a_2 > a_1^2/4$ and the roots μ_1, μ_2 are complex. Since a_1, a_2 are real, and μ_1, μ_2 are complex conjugates, and since $\mu_1\mu_2 = -a_2$, we can write the roots in the following (polar) form

$$\mu_1 = (-a_2)^{1/2}\, e^{i\,\theta}, \;\; \mu_2 = (-a_2)^{1/2}\, e^{-i\,\theta}$$

Using $\mu_1 + \mu_2 = a_1$, the angle θ is determined by

$$2(-a_2)^{1/2}\; cos\,\theta = a_1$$

giving

$$cos\,\theta = \frac{1}{2}a_1(-a_2)^{-1/2}. \tag{4.11}$$

Since the cosine function oscillates, the solution (4.11) ensures that the stationary process has an oscillatory form, and thus as described in Priestley (2004) is "damped periodic" or "pseudo-periodic". The inequality, $a_1^2 + 4a_2 < 0$, specifies a parabolic region within the triangular region of the (a_1, a_2) parameter plane (Figure 4.1, Shumway and Stoffer 2006).

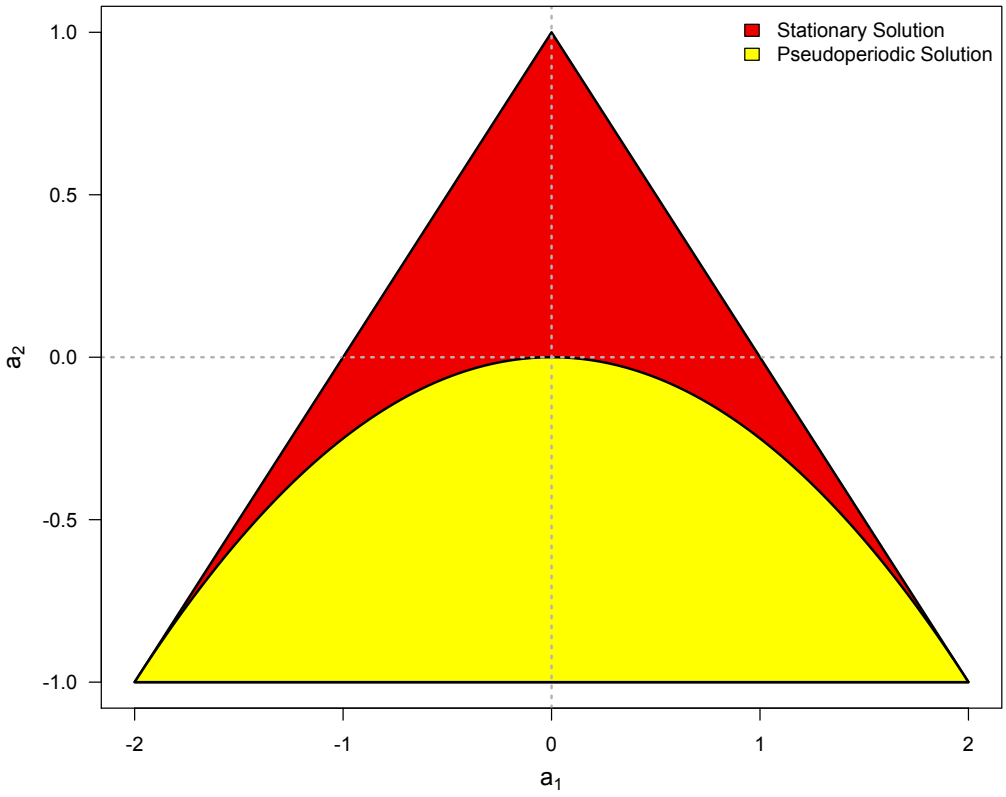

Figure 4.1: Parameter plane for the solution space of a set of movement parameters a_1 and a_2 that solve a set of stationary AR(2) difference equations.

The Behavioural Response Variable

The solution space for a_1, and a_2 identifies two distinct regions that characterize unique numerical properties (Figure 4.1). Our analysis proceeds by windowing the high density time series into 26 minute windows of 780 data points, and pre-identifying ocean-surface activity related to non-diving behaviour such as resting or surface-related swimming such as transiting between feeding areas. For purposes of the behavioural analyses, this group of surface behaviours was considered to be the same behavioural mode, as both behaviours are characterized by a lack of engagement in the immediate environment. This behaviour was called "Non-Diving", and typically had parameter solutions near zero indicating a purely random process with no correlation to previous time periods.

For windows that include dive behaviours we identified solutions that fall in the yellow region as repetitive diving reminiscent of an oscillating pendulum, and call the behaviour mode represented by solutions in this region "Active Diving". Parameter solutions that fall in the red area represent more intermittent diving characterized by a correlated random walk, and call the behaviour mode of solutions in this area "Exploratory Diving". Our analysis is based on the distinction defined by the numerical properties of the state space solution space, and ensures that we preserve the interpretability of the state space model for the analysis of seal behaviour (Priestly 2004).

We concatenated similar 26-minute windows of the behavioural solutions together into coherent "segments" of similar behaviour. This better relates the observed seal behaviour to the scale of the spatial covariate information and defines our response, $\boldsymbol{y}_{ij}$, as the observed behaviour of the i^{th} seal over the j^{th} segment of their at-sea voyage.

4.2 Observation (data) model : $P(\boldsymbol{y}_i|\boldsymbol{p}_i)$

Hierarchical Bayesian models account for the observational or data process via a data model. The data model is described in the first level of the hierarchy (Equation 4.3). This first level corresponds to the distribution of the data given the process of interest. The data are assumed to be generated by an observation process that depends on some environmental process. It is very possible that the outcome appears differently depending on which set of environmental factors are acting on it. This conditional relation is part of the environmental process, and can be formalized within the context of conditional probabilities. Here, we wish to specify a probability distribution for the outcome, $\boldsymbol{y}$, that is assumed to be summarized

(so far without error) by the ecological process, such that the uncertainty in this data model is assumed to be solely a result of the observation process (i.e. measurement error). This layer of hierarchy allows for the separation of error associated with the measurement of the outcome, and the errors associated with misspecification of the true underlying process. This distinction in errors is important in understanding the strength of hierarchical modelling.

One such data model is the multinomial logistic model, which is a generalized regression model that extends the linear regression model by allowing various nonnormal distributions for the response and error terms, and extends the more familiar logistic regression to allow the number of categorical responses to be greater than two (McCullagh and Nelder 1989). A density function for a set of multinomial random variables $\{Y\}$ with K possible outcomes $\{Y : Y^{(1)} = y^{(1)}, \ldots, Y^{(k)} = y^{(K)}\}$, and where $\sum_{k=1}^{K} y^{(k)} = \mathcal{K}$, has the form:

$$P(y^{(1)}, \ldots, y^{(K)}) = \frac{\mathcal{K}!}{\prod_{k=1}^{K} y^{(k)}!} \prod_{k=1}^{K} p^{(k)}. \tag{4.12}$$

This allows for multiple categorical outcomes such as multiple behavioural modes (eg. "Active Diving", "Exploratory Diving", and "Non-Diving"). In this application, each behavioural segment represents a variable-length section of time and space through which the fur seal swam through the ocean and engaged in one of a set of consistent behaviours. The observed behaviour class at each segment of each fur seal's at-sea track can then be modelled in relation to spatially and temporally linked covariate data.

4.2.1 A Multinomial Data Model for Estimating Fur Seal Behaviour

Let $X_{ij} = [X_{ij1}, \ldots, X_{ijp}]^T$ be the p-dimensional vector of linked covariate data characterizing the environment of the i^{th} animal over the j^{th} segment of the track in which the behaviour was observed. We encode the observed behaviour class by $k \in \{1, 2, \ldots, K\}$, a 0/1 valued vector of length of K, $\boldsymbol{y}_{ij} = (y_{ij}^{(1)}, y_{ij}^{(2)}, \ldots, y_{ij}^{(K)})$, where $y_{ij}^{(k)}$ is an indicator variable that takes the value 1 if the animal is observed in the k^{th} behaviour class and 0 otherwise, such that $\sum_{k=1}^{K} y_{ij}^{(k)} = 1$. The underlying (latent) probability vector describing the relative probabilities of each behaviour category $\{p_{ij}^{(1)}, p_{ij}^{(2)}, \ldots, p_{ij}^{(K)}\}$ can be linked to a set of covariates via the multinomial regression model. We model the logarithm of the ratio of the probability of each category relative to that of a baseline category. We selected the most commonly observed category (here, "Non-Diving") as our baseline category ($k = 1$),

following the advice in Agresti (1990). The multinomial logit linear model is then

$$log\left(\frac{P(\boldsymbol{y}_{ij}^{(k)}=1)}{P(\boldsymbol{y}_{ij}^{(1)}=0)}\right)=\boldsymbol{X}_{ij}\boldsymbol{\beta}_i^{(k)}, \tag{4.13}$$

where $\boldsymbol{\beta}_i^{(k)}$ is the parameter vector of regression coefficients corresponding to class k; i.e. $\boldsymbol{\beta}_i^{(k)}=\{\beta_{i1}^{(k)},\ldots,\beta_{ip}^{(k)}\}^T$. The probability that y_{ij} belongs to class k is written as

$$P(y_{ij}^{(k)}=1|\boldsymbol{X}_{ij},\boldsymbol{\beta}_i)=\frac{exp(\boldsymbol{X}_{ij}\boldsymbol{\beta}_i^{(k)})}{\sum_{m=1}^{K}exp(\boldsymbol{X}_{ij}\boldsymbol{\beta}_i^{(m)})}, \tag{4.14}$$

where the joint likelihood for all observed data of the i^{th} animal is

$$P(\boldsymbol{y}_i|\boldsymbol{X}_i,\boldsymbol{\beta}_i)=\prod_j\prod_{k=1}^{K}\left(\frac{exp(\boldsymbol{X}_{ij}\boldsymbol{\beta}_i^{(k)})}{\sum_{m=1}^{K}exp(\boldsymbol{X}_{ij}\boldsymbol{\beta}_i^{(m)})}\right)^{y_{ij}^{(k)}}; \tag{4.15}$$

(Gelman et al. 2004, p. 430). Equation (4.14) defines the probability of animal i, over track segment j engaging in behaviour k, as $p_{ij}^{(k)}$, and we have assumed that $y_{ij}^{(k)}\sim Multinomial(p_{ij}^{(k)})$, and the parameters that describe the relation between $\boldsymbol{y}_i$ and $\boldsymbol{p}_i$, as $\boldsymbol{\beta}_i^{(k)}$.

Because of the condition $\sum_{k=1}^{K}P(y_{ij}^{(k)}=1|X_{ij},\beta_i)=1$, the parameter vector for one of the classes need not be estimated, and we can set one of the parameter vectors $\boldsymbol{\beta}_i^{(1)}=\mathbf{0}$. The joint likelihood for the multinomial data model can then be rearranged as

$$P(\boldsymbol{y}_i|\boldsymbol{X}_i,\boldsymbol{\beta}_i)=exp\left\{\sum_{k=2}^{K}\left(\boldsymbol{\beta}_i^{(k)}\left[\sum_j y_{ij}^{(k)}\boldsymbol{X}_{ij}\right]\right)-\sum_j log\left(1+\sum_{k=2}^{K}e^{\boldsymbol{X}_{ij}\boldsymbol{\beta}_i^{(k)}}\right)\right\}. \tag{4.16}$$

For simplicity, we refer to the set of $K-1$ m-dimensional regression coefficients $\{\beta_{i1}^{(2)},\ldots,\beta_{im}^{(2)},\ldots,\beta_{i1}^{(K)},\ldots,\beta_{im}^{(K)}\}$ for the i^{th} animal as $\boldsymbol{\beta}_i$, with likelihood (Equation 4.16) written

$$P(\boldsymbol{y}_i|\boldsymbol{X}_i,\boldsymbol{\beta}_i)=exp\left\{\sum_j \boldsymbol{y}_{ij}\boldsymbol{\beta}_i\boldsymbol{X}_{ij}-\sum_j log\left(1+e^{\boldsymbol{X}_{ij}\boldsymbol{\beta}_i}\right)\right\}. \tag{4.17}$$

For this study, we assumed that each segment of track of each animal follows the same data model (i.e. Equation 4.14), while allowing for changes of resolution and alignment and measurement-error characteristics between animals.

4.3 Process Model : $P(\boldsymbol{p}_i|\boldsymbol{\theta}_i)$

The process model is central to the hierarchical modelling framework, and here, represents the 2^{nd} level of our hierarchy i.e. Equation 4.4. The parameter vector governs the process that is of biological interest and is rarely observed. We have assumed until now (i.e. in Section 4.2) that the (environmental) process was deterministic and errors were only in the observation of $\boldsymbol{y}_i$, but not in the parameters that describe the environmental process $\boldsymbol{\beta}_i$. In the conditional modelling framework of the hierarchical Bayesian model (described in Section 4.1.2), we include the uncertainty of the process model.

The process model is composed of explicit and implicit processes (Royle and Dorazio 2008). A model can be explicit in the sense that there is biological understanding of the environmental process that generated the animal's behavioural response, and these are described by the process parameters $\boldsymbol{\beta}_i$. An implicit process model is one that includes a "random effect" that serves as a surrogate for a real environmental process, but one that is difficult to characterize as it is poorly informed about the data (or not at all). One example of an implicit process is the uncertainty in scale of measurement of an animal's environment. This random effect is an implicit process because the "process" lacks an explicit biological interpretation. Getting the explicit form of the process model right implies we have correctly identified the causal mechanism for our data so the model results are scientifically meaningful and interpretable (Cressie et al. 2009).

4.3.1 Error-in-Covariate models

An important aspect of process error problems is the difficulty in accurately estimating parameters of interest when there is error contained within the covariate data (Stephens and Dellaportas 1992). Standard regression models assume that all covariates have been measured exactly, i.e. without error; as such, those models account only for errors in the dependent variables, or responses. In the case when some regressors have been measured with errors, estimation based on the standard assumption of no error in the covariates leads to biased and inaccurate inference about the true underlying response-covariate associations (Gustafson 2004). In particular, the parameter estimates do not tend to the true values even in very large samples, and for simple linear regression the effect is an underestimate of the coefficient, known as the *attenuation bias* (Becker 1986). In non-linear models, such as the multinomial logit model, the direction of the bias is more complicated (Schennach 2004)

and less predictable, and hence more important to model explicitly.

Many of the spatial covariates remain consistent over large sections (wind speed, surface temperature), while others change at finer scales (groundfish catch, primary productivity). Thus depending on the time spent in a behaviour mode, there may be a distribution of values over that portion of the track. In a previous section (Section 4.2.1), all measured covariates for the i^{th} fur seal were represented by $\boldsymbol{X}_i$. Here we distinguish between covariates measured with and without error, as well we redefine $\boldsymbol{X}_i$ to be a covariate that we don't directly measure, but instead infer from observing a related variable $\boldsymbol{W}_i$. By modelling this error-in-covariates phenomenon, we are assuming an implicit random effect model that informs us about the true data $\boldsymbol{X}_i$ through an error prone measure $\boldsymbol{W}_i$.

For clarification we provide the following guide to error-in-covariate notation for the i^{th} fur seal (duplicated in Appendix C), which follows the notation of Carroll et al. (2006):

$\boldsymbol{p}_i$	The usual multinomial probabilities
$\boldsymbol{W}_i$	Observations related to $\boldsymbol{X}_i$, measured with error
$\boldsymbol{X}_i$	Unobserved "true" covariate(s)
$\boldsymbol{U}_i$	Measurement error in an error model
$\boldsymbol{Z}_i$	Covariate(s) measured without error.

We consider a multinomial regression model to infer behavioural patterns for the i^{th} fur seal from a set of covariates, one (or more) of which is a covariate measured with error. In other words, we want to formulate a statistical model that is fit to well-defined but unobservable covariate(s), $\boldsymbol{X}_i$, using information from $\boldsymbol{W}_i$, that are less than perfectly correlated with $\boldsymbol{X}_i$, and a set of covariates $\boldsymbol{Z}_i$ measured without error. We rewrite Equation (4.13) for the j^{th} segment of the i^{th} fur seal's track as

$$log\left(\frac{\boldsymbol{p}_{ij}^{(k)}}{\boldsymbol{p}_{ij}^{(1)}}\right) = \boldsymbol{Z}_{ij}\boldsymbol{\beta}_{iZ} + \boldsymbol{X}_{ij}\boldsymbol{\beta}_{iX}. \tag{4.18}$$

Here, the goal is to obtain unbiased estimates of parameters $\boldsymbol{\beta}_{iZ}$ and $\boldsymbol{\beta}_{iX}$ indirectly by fitting a model for $logit\ \boldsymbol{p}_{ij}^{(k)}$ in terms of $(\boldsymbol{Z}_{ij}, \boldsymbol{W}_{ij})$. The problem is that the parameters of the regression of $logit\ \boldsymbol{p}_{ij}^{(k)}$ on $(\boldsymbol{Z}_{ij}, \boldsymbol{W}_{ij})$ are different from those of $logit\ \boldsymbol{p}_{ij}^{(k)}$ on $(\boldsymbol{Z}_{ij}, \boldsymbol{X}_{ij})$, and substituting $\boldsymbol{W}_{ij}$ for $\boldsymbol{X}_{ij}$, while making no adjustments in the usual fitting methods for this substitution, leads to biased estimates (Gustafson 2004, Carroll et al. 2006).

In assessing measurement error, careful attention must be given to the type and nature of the error, and the sources of data that allow modeling of this error (Carroll et al. 2006).

In classical error-in-covariates models, the observed measure, $\boldsymbol{W}_{ij}$, equals the true measure plus some measurement error, i.e.

$$\boldsymbol{W}_{ij} = \boldsymbol{X}_{ij} + \boldsymbol{U}_{ij}. \tag{4.19}$$

This formulation assumes the variability of the observed measure will be greater than the "true" measure due to error in the measuring device or context.

In our context, the major source of covariate "error" is the fact that the environmental field is divided up into segments, for which there is one measure of the covariate per segment, and a single measure is an inadequate measure of the range of values experienced during that segment of at-sea track. For example, there could be diurnal changes at the same location, vertical differences, or finer-scale horizontal differences than the measurement resolution, or a combination of these types of errors.

If the observed measure, $\boldsymbol{W}_{ij}$, is an index of average abundance for the segment of track without substantial measurement error, then we rearrange the classical error model to reflect where the error lies (i.e. the truth is more variable than the estimator), and instead form an error model as simple as Equation (4.19) with $\boldsymbol{W}_{ij}$ and $\boldsymbol{X}_{ij}$ reversed, i.e.

$$\boldsymbol{X}_{ij} = \boldsymbol{W}_{ij} + \boldsymbol{U}_{ij}. \tag{4.20}$$

We assume $\boldsymbol{W}_{ij}$ is unbiased for $\boldsymbol{X}_{ij}$, and $\boldsymbol{U}_{ij}$ is a normal, independent and identically distributed random variable according to $\sim \mathcal{N}(\boldsymbol{0}, \sigma^2_{iU})$. The random variable $\boldsymbol{U}_{ij}$ is parameterized by a variance term σ^2_{iU} that describes the covariate uncertainty for the i^{th} fur seal, such that we have presumed a process model with normally-distributed random effects (or process error). In this formulation, it is assumed that there is more error in the true $\boldsymbol{X}_{ij}$ than in the observed data $\boldsymbol{W}_{ij}$, i.e. $Var(\boldsymbol{X}_i) = Var(\boldsymbol{W}_i) + \sigma^2_{iU}$. We assume the covariate error is independent from the "observed" value of the covariate, and we assume that there is no bias in the measurement, i.e. $E(\boldsymbol{U}_{ij}|\boldsymbol{W}_{ij}, \boldsymbol{Z}_{ij}) = \boldsymbol{0}$ and $Var(\boldsymbol{U}_{ij}|\boldsymbol{W}_{ij}, \boldsymbol{Z}_{ij}) = \sigma^2_{iU}$. We rewrite Equation (4.18) as

$$log\left(\frac{\boldsymbol{p}^{(k)}_{ij}}{\boldsymbol{p}^{(1)}_{ij}}\right) = \boldsymbol{Z}_{ij}\boldsymbol{\beta}_{iZ} + \boldsymbol{W}_{ij}\boldsymbol{\beta}_{iX} + \boldsymbol{U}_{ij} \tag{4.21}$$

and the process model is then

$$P(\boldsymbol{p}^{(k)}_i|\boldsymbol{Z}_i, \boldsymbol{X}_i, \boldsymbol{\beta}^{(k)}_i) \sim \mathcal{N}\left(\left\{ \frac{exp(\boldsymbol{Z}_{ij}\boldsymbol{\beta}^{(k)}_{iZ} + \boldsymbol{W}_{ij}\boldsymbol{\beta}^{(k)}_{iX})}{\sum_{m=1}^{K} exp(\boldsymbol{Z}_{ij}\boldsymbol{\beta}^{(m)}_{iZ} + \boldsymbol{W}_{ij}\boldsymbol{\beta}^{(m)}_{iX})} \right\}^{y^{(k)}_{ij}}, \sigma^2_{iU} \right). \tag{4.22}$$

Errors of the type found in Equation (4.20) are called Berkson errors (Berkson 1950), and are typically chosen when all observations (along a segment of track) within a group (or individual) are given the same error-prone covariate. An example to illustrate the ideas of modelling Berkson error-in-covariates is to consider a single covariate that is measured with error. Consider the case of a single (j^{th}) segment for the i^{th} fur seal of the covariate "$groundfish\ catch$", then

$\boldsymbol{X}_{ij}$ Unobserved groundfish abundance experienced by the fur seal
$\boldsymbol{W}_{ij}$ Average groundfish catch in the Alaska commercial fishery

In our application, fur seals swimming through a particular spatial grid square are all given the same groundfish measure for that segment of track, except that the true groundfish experienced by the individual is particular to the individual herself (and unlikely to be exactly the same as the value assigned to that grid square). The Berkson model assumes that the true groundfish catch is equal to the estimated groundfish catch plus random error, so that the true groundfish catch has more variability than the estimated groundfish catch as it is likely the seal experiences a range of groundfish densities throughout her behaviour segment.

Bayesian Error-in-Covariates Model

To incorporate the uncertainty in the covariates within a Bayesian framework, we treat $\boldsymbol{X}_i$ as a latent (unobservable) variable, and model its distributional form using one or more observed variables assumed to contain information about $\boldsymbol{X}_i$, i.e. $\boldsymbol{W}_i$ (Richardson and Gilks 1993, Carroll et al. 2006).

For the i^{th} fur seal, we include this uncertainty in the environmental covariate in the lower-level process models by estimating the distribution of the true predictor $\boldsymbol{X}_i$ using the observed variable $\boldsymbol{W}_i$, and treat the covariate model (Equation 4.20) as the prior for $\boldsymbol{X}_i$. Uncertainty in the covariate error is incorporated into the process model and not implemented into higher level (parameter) models such that the measurement error in a covariate is fit to each animal, independent of other animal's data. The variance terms, σ^2_{iU} for each fur seal, are then treated as nuisance parameters, and fit without shrinkage.

The steps to adding error-in-covariates to a Bayesian multinomial logit regression are as follows (adapted from Carroll et al. 2006, Gimenez et al. 2009):

1. Identify covariates with measurement errors, and select the error model (e.g. Berkson).

2. Identify the likelihood. The likelihood function is exactly the same as (4.17), except the likelihood has covariates observed without error $\boldsymbol{Z}_i$, and unobserved covariates $\boldsymbol{X}_i$ that are treated as if they were observed with a normal, identically distributed error distribution.

3. Select priors for all unknowns, where the unknowns in this model are $(\boldsymbol{X}_{i1}, \ldots, \boldsymbol{X}_{il}, \ldots, \boldsymbol{X}_{ip}, \sigma^2_{iU}, \boldsymbol{\beta}_{iZ}, \boldsymbol{\beta}_{iX})$.

4. Compute complete conditionals for all unknowns from the following joint posterior density (see Appendix B),

$$P(\boldsymbol{\beta}_{iZ}, \boldsymbol{\beta}_{iX}, \boldsymbol{X}_i, \sigma^2_{iU} | \boldsymbol{y}_i, \boldsymbol{Z}_i, \boldsymbol{W}_i, \cdot) \propto P(\boldsymbol{y}_i | \boldsymbol{\beta}_{iZ}, \boldsymbol{\beta}_{iX}, \boldsymbol{Z}_i, \boldsymbol{X}_i, \boldsymbol{W}_i, \sigma^2_{iU}) P(\boldsymbol{X}_i) \quad \times$$
$$P(\sigma^2_{iU}) P(\boldsymbol{\beta}_{iZ}, \boldsymbol{\beta}_{iX}) \qquad (4.23)$$

5. Use conditional models to alternate between sample-based solutions for $\boldsymbol{X}_i$ and $\boldsymbol{\beta}_{iZ}, \boldsymbol{\beta}_{iX}$ as follows:

   (a) Generate a completely updated $\boldsymbol{X}_i$ using the current $\boldsymbol{\beta}_{iZ}, \boldsymbol{\beta}_{iX}$ parameter values, as well as the observed data $(\boldsymbol{y}_i, \boldsymbol{Z}_i, \boldsymbol{W}_i)$ (noting that we condition on $\boldsymbol{y}_i$ as well as $\boldsymbol{Z}_i$ and $\boldsymbol{W}_i$),

$$P(\boldsymbol{X}_i | \cdot) \propto P(\boldsymbol{y}_i | \boldsymbol{\beta}_{iZ}, \boldsymbol{\beta}_{iX}, \boldsymbol{X}_i, \boldsymbol{Z}_i) P(\boldsymbol{X}_i | \boldsymbol{W}_i, \sigma^2_{iU}) P(\sigma^2_{iU})$$

   (b) generate an updated set of $\boldsymbol{\beta}_{iZ}, \boldsymbol{\beta}_{iX}$ using the completed data set of both $\boldsymbol{X}_i$, now filled-in from step (a), and $\boldsymbol{Z}_i$.

$$P(\boldsymbol{\beta}_{iZ}, \boldsymbol{\beta}_{iX} | \cdot) \propto P(\boldsymbol{y}_i | \boldsymbol{\beta}_{iZ}, \boldsymbol{\beta}_{iX}, \boldsymbol{X}_i, \boldsymbol{Z}_i) P(\boldsymbol{\beta}_{iZ}, \boldsymbol{\beta}_{iX})$$

Henceforth, to simplify notation, we write $\boldsymbol{\beta}_i$ to represent the regression parameter set $\{\boldsymbol{\beta}_{iZ}, \boldsymbol{\beta}_{iX}\}$ for the parameters specific to the i^{th} fur seal unless there is uncertainty in notation in which case we use explicit notation. We write $\boldsymbol{\beta}$ for the collective set of regression parameters for all fur seals, and write $\boldsymbol{\theta}$ to represent the collective set of unknowns $\{\boldsymbol{X}_{i1}, \ldots, \boldsymbol{X}_{il}, \ldots, \boldsymbol{X}_{ip}, \sigma^2_{iU}, \boldsymbol{\beta}_{iZ}, \boldsymbol{\beta}_{iX}\}$.

Note that the above combination of *Observation* and *Process* models is similar in essence to a state space model in that a process model is updated with data to estimate a posterior density. This target posterior is then a scaled product of the likelihood function of $\boldsymbol{y}_i$ and the

process model $\boldsymbol{p}_i$. We assume that conditional on the covariates and associated regression parameters, all data observations are independent, so the joint density of the data model and process model can be combined and written

$$P(\boldsymbol{y}|\boldsymbol{X}, \boldsymbol{Z}, \boldsymbol{\beta}_i, \boldsymbol{\sigma^2})$$

4.4 Parameter Model : $P(\boldsymbol{\theta}|\boldsymbol{\psi})$

One of the great challenges in understanding how adaptive fur seal behaviour is to changes in the environment, is to understand the diversity of behaviour across individuals. To understand fur seal behaviour in a framework that targets population heterogeneity requires both estimates of regression parameters, but also a characterization of the inter-seal variability parameters. Including a parameter model is a means to formally account for variation between individual fur seals. In this "random effects" perspective, the tagged fur seals (more generally called *units*) are viewed as a random sample from a larger population to which we are to extend our inference.

Priors for unit-level regression parameters

Frequently there is some uncertainty over what the parameters of the unit-level priors should be, especially when we wish to capture the complicated structure of a dataset with multiple observations taken on a single fur seal (or unit). One naïve way of selecting priors for $\boldsymbol{\beta}$ is to assume the prior information for one fur seal is unrelated in any way to the information of any other fur seal; i.e., selecting an independent set of priors for each fur seal. This leaves us with a set of unrelated posterior inferences relating each animal's behaviour to the environmental process. An alternate way to select priors for $\boldsymbol{\beta}$ is to relate regression equations through correlations between the unit-level regression coefficient vectors. If the prior is structured as a series of conditional distributions that assume that unit-level regression parameters are conditionally independent, but share the same prior distribution, then this is the framework of a parameter model (Ntzoufras 2009; Cressie et al. 2009).

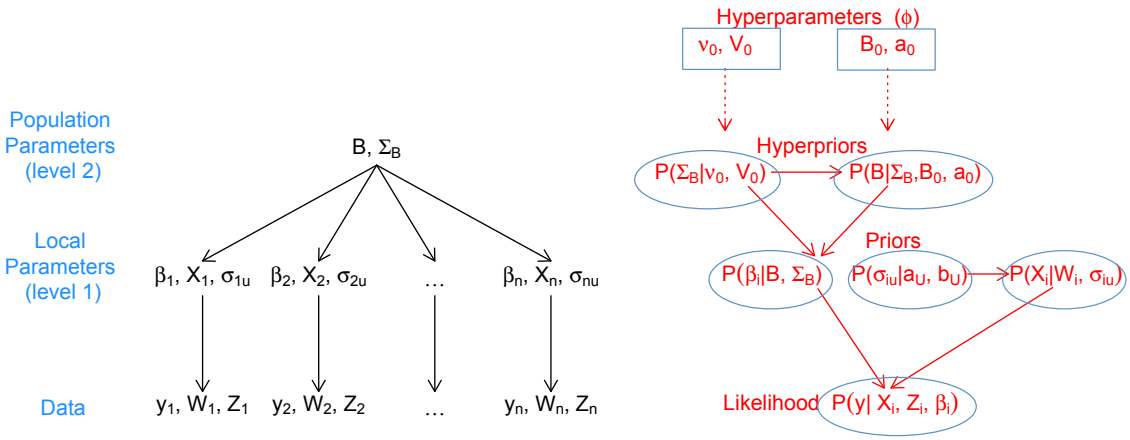

Figure 4.2: Graphical representation of a 2-stage Bayesian hierarchical model showing the hierarchy induced by sampling n random, independent animals from a population. The right side depicts the directed acyclic graph in which the square node refers to the constant (hyper) parameters, and the oval nodes refer to stochastic components of the model. By fitting a set of conditional models we have accounted for the uncertainty in the process governing fur seal behaviour in the higher level model, and the uncertainty in the data-generating mechanism in the lower-level model (Calder et al. 2003, Cressie et al. 2009).

By modelling the parameter model as a separate level in a hierarchical model, we are building a third layer ("level 2") that acknowledges a different source of randomness (Cressie et al. 2009). The lower-level specification of the model is specific to the individual animal, and expresses the uncertainty we have about that animal's response to the ecological process we'd like to understand. Therefore, we can think of the set of regression parameters specific to a single animal (e.g. $\boldsymbol{\beta}_i$) as a sample from a distribution of possible values (i.e. random coefficients). By adopting a two-stage prior distribution for $\boldsymbol{\beta}$ that assigns the same hyper-parameters $\{\boldsymbol{\phi}\}$ on across-unit priors $\{\boldsymbol{\psi}\}$, we are formally acknowledging our uncertainty over what the priors are for the lower-level parameters, as well as linking all the lower-level models. By linking regression models via a shared prior, we can "borrow strength" between animals (Ntzoufras 2009), and the posterior of each unit-level $\boldsymbol{\beta}_i$ is then a weighted mean of the corresponding animal's regression and the overall population effect $\boldsymbol{B}$.

The population-level parameters $\boldsymbol{\psi} = \{\boldsymbol{B}, \boldsymbol{\Sigma}_\beta\}$ are found by writing the joint Bayesian posterior distribution of $\{\boldsymbol{\beta}_1, \ldots, \boldsymbol{\beta}_n, \boldsymbol{B}, \boldsymbol{\Sigma}_\beta\}$, and decomposing this into separate conditional distributions for the population parameters, and the local parameters updated by the data (e.g., Wikle et al. 2003). The joint prior distribution is

$$P(\boldsymbol{B}, \boldsymbol{\Sigma}_\beta, \boldsymbol{\beta}_1, \ldots, \boldsymbol{\beta}_p) \;=\; \prod_i \Big(P(\boldsymbol{\beta}_i | \, \boldsymbol{B}, \boldsymbol{\Sigma}_\beta) \Big) P(\boldsymbol{B}, \boldsymbol{\Sigma}_\beta),$$

and the joint posterior distribution is

$$P(\boldsymbol{B}, \boldsymbol{\Sigma}_\beta, \boldsymbol{\beta}_1, \ldots, \boldsymbol{\beta}_p | \boldsymbol{y}, \cdot) \;\propto\; \prod_i \Big(P(\boldsymbol{y}_i | \boldsymbol{\beta}_i, \boldsymbol{B}, \boldsymbol{\Sigma}_\beta, \cdot) \times P(\boldsymbol{\beta}_i | \, \boldsymbol{B}, \boldsymbol{\Sigma}_\beta) \Big) P(\boldsymbol{B}, \boldsymbol{\Sigma}_\beta) \tag{4.24}$$

$$= \prod_i \Big(P(\boldsymbol{y}_i | \boldsymbol{\beta}_i, \cdot) P(\boldsymbol{\beta}_i | \, \boldsymbol{B}, \boldsymbol{\Sigma}_\beta) \Big) P(\boldsymbol{B}, \boldsymbol{\Sigma}_\beta) \tag{4.25}$$

with the simplification of Equation (4.25) from (4.24) because the process model likelihood depends only on the process parameters $\boldsymbol{\beta}$, and the prior distributions of the parameter model, $P(\boldsymbol{B}, \boldsymbol{\Sigma}_\beta)$, affect $\boldsymbol{y_i}$ only through $\boldsymbol{\beta}_i$ (Gelman et al. 2004, p 124). Therefore from these 1^{st}-stage conditional densities, we produce an unconditional higher-order likelihood. It is possible that this higher-level likelihood has covariates that describe the lower-level units (for our example, this could be seal weight, or age), and will be denoted as $\boldsymbol{\mathcal{X}}$ in the following likelihood and density functions.

In selecting the two-stage prior distribution for the unit-level parameters $\boldsymbol{\beta}_i$, the first stage priors were $\sim \mathcal{N}(\boldsymbol{\mathcal{X}B}, \boldsymbol{\Sigma}_\beta)$ and because of this assumption at the first stage, we can use

normally distributed conjugate priors at the higher (2^{nd}) levels. The posterior estimates of $\boldsymbol{\psi}$, then form the higher-level parameters that describe the overall population effect and our uncertainty in the environmental process. The lower-level parameters $\boldsymbol{\beta}$ are now interpreted as a random sample from a multivariate normal distribution with expectation $\boldsymbol{B}$ and variance $\boldsymbol{\Sigma}_\beta$. The joint state space of all the unknowns in our model is proportional to the following

$$P(\boldsymbol{\beta}, \boldsymbol{X}, \sigma^2_U, \boldsymbol{B}, \boldsymbol{\Sigma}_\beta | \boldsymbol{y}, \boldsymbol{Z}, \boldsymbol{\mathcal{X}}) \propto \prod_i \Bigg(P(\boldsymbol{y}_i|\boldsymbol{\beta}_i, \boldsymbol{Z}_i, \boldsymbol{X}_i, \boldsymbol{\mathcal{X}}) \times P(\boldsymbol{X}_i|\boldsymbol{W}_i, \sigma^2_{iU}) \; P(\boldsymbol{\beta}_i|\boldsymbol{B}, \boldsymbol{\Sigma}_\beta, \boldsymbol{\mathcal{X}}) \; P(\sigma^2_{iU}) \Bigg) \times P(\boldsymbol{B}, \boldsymbol{\Sigma}_\beta | \boldsymbol{\mathcal{X}}, \boldsymbol{\phi}) \tag{4.26}$$

For inferential statements about parameters separately, we need the marginal posterior distribution for each unknown. For example, a point estimate or credible interval for $\boldsymbol{\beta}_i$ can be obtained by integrating over all other unknowns of Equation (4.26)

$$P(\boldsymbol{\beta}_i|\boldsymbol{y}, \boldsymbol{Z}, \boldsymbol{\mathcal{X}}) \propto \int_{\boldsymbol{\beta}_{-i}} \int_{\boldsymbol{X}} \int_{\sigma^2_U} \int_{\boldsymbol{B}} \int_{\boldsymbol{\Sigma}_\beta} \prod_i \Bigg(P(\boldsymbol{y}_i|\boldsymbol{\beta}_i, \boldsymbol{Z}_i, \boldsymbol{X}_i, \boldsymbol{\mathcal{X}}) \times P(\boldsymbol{X}_i|\boldsymbol{W}_i, \sigma^2_{iU}) \; P(\boldsymbol{\beta}_i|\boldsymbol{B}, \boldsymbol{\Sigma}_\beta, \boldsymbol{\mathcal{X}}) \; P(\sigma^2_{iU}) \Bigg) \times P(\boldsymbol{B}, \boldsymbol{\Sigma}_\beta | \boldsymbol{\mathcal{X}}, \boldsymbol{\phi}) \; d\boldsymbol{\beta}_{-i} \; d\boldsymbol{X} \; d\sigma^2_U \; d\boldsymbol{B} \; d\boldsymbol{\Sigma}_\beta. \tag{4.27}$$

There is no closed form for the integrations of the marginal posterior distribution of Equation (4.27); likewise for the marginals of the other parameters in Equation (4.26). Instead a set of conditional posterior distributions can be derived for each parameter of (4.26) from which we set up a composition sampling method to draw samples from each conditional posterior distribution iteratively. The result is that the collection of samples from the set of conditional posteriors derived from the Bayesian hierarchical multinomial model with error-in-covariates of Equation (4.26) is a sample from its joint posterior distribution. This is discussed further in Section (4.5).

We write out the joint distribution of Equation (4.26) for the parameters of interest as the following sequence of conditional posterior distributions (All symbols subscripted by 0, indicate dirac delta measures (Section 2.2.3) of parameters describing hyper-prior

distributions, i.e. $\boldsymbol{\phi} = \boldsymbol{V}_0, \boldsymbol{B}_0, \nu_0, a_0$; derived full conditionals appear in Appendix B).

$$P(\boldsymbol{\beta}_i | \, \boldsymbol{y}_i, \boldsymbol{X}_i, \boldsymbol{Z}_i, \boldsymbol{\mathcal{X}}, \boldsymbol{B}, \boldsymbol{\Sigma}_\beta) \tag{4.28}$$

$$P(\boldsymbol{X}_i | \, \boldsymbol{y}_i, \boldsymbol{W}_i, \boldsymbol{Z}_i, \boldsymbol{\beta}_i, \sigma^2_{iU}) \tag{4.29}$$

$$P(\sigma^2_{iU} | \boldsymbol{X}_i, \boldsymbol{W}_i, a_U, b_U) \tag{4.30}$$

$$P(\boldsymbol{\Sigma}_\beta | \boldsymbol{\beta}_i, \boldsymbol{\mathcal{X}}, \boldsymbol{V}_0, \boldsymbol{B}_0, \nu_0, a_0) \tag{4.31}$$

$$P(\boldsymbol{B} | \boldsymbol{\Sigma}_\beta, \boldsymbol{\mathcal{X}}, \boldsymbol{V}_0, \boldsymbol{B}_0, \nu_0, a_0, \boldsymbol{\beta}_i). \tag{4.32}$$

.

4.5 Algorithm for Bayesian Hierarchical Error-in-Covariates

4.5.1 Monte Carlo Methods

The joint posterior distribution for all the unknowns in the model has been written in explicit form in Equation (4.26), but inference around parameters in this joint distribution depend on multiple integrations for each marginal posterior as in Equation (4.27), and an analytical solution (in most cases) is not possible. Therefore, we turn to approximate solutions that converge in distribution to the true distribution. Markov Chain Monte Carlo (MCMC) is a general iterative algorithm that draws a sequence of random samples from a proposal distribution following a Markov assumption, and bases inference about unknown variables on the random draws from this distribution. In Section (2.2.3 ; Equations 2.9 through 2.11), we introduced Monte Carlo (MC) integration as a way of using a random sampler to compute an integral, $\mathcal{I}$, and showed how MC integration can be used to approximate the posterior (or marginal posterior) distributions of a Bayesian analysis.

One problem with applying Monte Carlo integration is in obtaining samples from some complex probability distribution, i.e. $g(x)$ in Equation (2.9), in particular when $g(x)$ is a high-dimensional probability density function as in Equation (4.26). Attempts to solve this problem are the roots of MCMC methods. As before, we let the state space refer to the range of possible values for a random variable x, and describe a random variable of a Markov process at iteration t, $x^{(t)}$, to only depend on its previous state at iteration $t-1$, $x^{(t-1)}$; i.e.,

$$x^{(t)} \sim f(\cdot | x^{(t-1)}).$$

When there is more than one random variable in the state space at each Markov iteration, we can sequentially update the state components of the (now) vector valued $\boldsymbol{x} = \{x_1, \ldots, x_m\}$ in turn by drawing from the full conditionals of the components in $\boldsymbol{x}$. For example, at iteration t in which we've already sampled from $x_1, \ldots, x_{j-1}$, we update the j^{th} component of the state vector $\boldsymbol{x}^{(t)}$ as follows:

$$x_j^{(t)} \sim f_{x_j|x_{j-1}}(\cdot|x_1^{(t)}, \ldots, x_{j-1}^{(t)}, x_{j+1}^{(t-1)}, \ldots, x_m^{(t-1)}) \tag{4.33}$$

Thus, one iteration of a Markov chain at iteration t involves updating components of the joint distribution described in Equation (4.26), in turn by sequentially drawing from the set of full conditionals of Equations (4.28) through (4.31). Each proposal is accepted or rejected according to a criterion called the acceptance ratio in the Metropolis-Hastings algorithm (Metropolis et al. 1953; Hastings 1970), or is always accepted if the samples are drawn directly from the target distribution as in Gibbs sampling (Geman and Geman 1984, Gelfand and Smith 1990). We generate an MCMC by repeatedly drawing random samples from the joint state space using proposal distributions based on the previous iteration's vector of values, and accepting (or rejecting) those proposal values. A MCMC refers to a sequence of random samples of the (multivariate) state, generated by a Markov process, and these random samples constitute a random walk over the (joint) posterior distribution. MCMC differs from standard Monte Carlo integration however, in that the samples are dependent, since they are realized paths of a Markov chain. Once thinned appropriately (to remove the within-chain dependence), the retained random samples are then equivalent to integrating the joint density over the data to obtain a sample from the posterior distribution for the model unknowns.

The random samples of the state after a large number of steps along the Markov Chain are then used as samples from the desired distribution, and these samples of the state values are spread out according to the posterior (target) probability density function. At equilibrium, an effective MCMC is one that has converged to be irreducibile (all states communicate with all other states), reversible (satisfies the condition of detailed balance) and aperiodic (cycles through all states with no fixed number of steps before it returns to its current value; Wasserman 2004). Once the MCMC has reached the stationary target distribution and can be shown to be irreducible, reversible and aperiodic, the inferences are based on the properties of random samples of the state vector.

Hierarchical statistical modelling leads to computationally intensive inference (Gilks et

al. 1995). Recent increased use of these kinds of models has coincided with the revolution in Bayesian computation and further development of MCMC simulation approaches (Wikle et al. 2003). MCMC methods can be used in Bayesian hierarchical modelling. These sampling based methodologies make it relatively efficient to sample the solution space, and allow for the extension to ever-more-complicated modelling scenarios (Gelfand and Smith 1990). Below is a discussion of the algorithm that we built for sampling from the joint posterior of a hierarchical error-in-covariates multinomial model.

4.5.2 Lower-level (first-stage/random effects) model : A Random-Walk Metropolis-Hastings step for $\boldsymbol{\beta}_{iX}, \boldsymbol{\beta}_{iZ}$

We set up an MCMC sampling scheme for sampling from the joint state space of Equation (4.26), by constructing the conditional distributions of (4.28 through 4.31). To begin, we construct the full conditional distribution of the lower-level model (4.28), as follows:

$$\begin{aligned} P(\boldsymbol{\beta}_i|\boldsymbol{y}_i, \boldsymbol{X}_i, \boldsymbol{Z}_i, \boldsymbol{B}, \boldsymbol{\Sigma}_\beta) &\propto P(\boldsymbol{y}_i|\boldsymbol{Z}_i, \boldsymbol{X}_i, \boldsymbol{\beta}_i) P(\boldsymbol{\beta}_i|\boldsymbol{\mathcal{X}}, \boldsymbol{B}, \boldsymbol{\Sigma}_\beta) \\ &\propto exp\left\{ \sum_j \boldsymbol{y}_{ij} \left(\boldsymbol{\beta}_{iZ}\boldsymbol{Z}_{ij} + \boldsymbol{\beta}_{iX}\boldsymbol{X}_{ij}\right) - \right. \\ &\quad \left. \sum_j log\left(1 + e^{\boldsymbol{\beta}_{iZ}\boldsymbol{Z}_{ij} + \boldsymbol{\beta}_{iX}\boldsymbol{X}_{ij}}\right) \right\} \times \\ &\quad |\boldsymbol{\Sigma}_\beta|^{-n/2} exp\left\{ -\frac{1}{2} tr(\boldsymbol{\beta}_i - \boldsymbol{\mathcal{X}}\boldsymbol{B}) \boldsymbol{\Sigma}_\beta^{-1} (\boldsymbol{\beta}_i - \boldsymbol{\mathcal{X}}\boldsymbol{B}) \right\} \end{aligned} \tag{4.34}$$

We assume that each of the m components of $\boldsymbol{\beta}_i$ is normally distributed, and the prior distribution of the $\boldsymbol{\beta}$ is then a multivariate normal distribution ($\mathcal{MVN}$). We use the posterior estimates of these higher level parameters $\{\boldsymbol{B}, \boldsymbol{\Sigma}_\beta\}$ to inform the prior parameters that describe the distribution of the lower-level regression parameters $\boldsymbol{\beta}_i$, i.e. $P(\boldsymbol{\beta}_i) \sim \mathcal{MVN}(\boldsymbol{\mathcal{X}}\boldsymbol{B}, \boldsymbol{\Sigma}_\beta)$.

Since the likelihood of (4.34) is of multinomial form, the normal distribution is not a convenient conjugate prior for $\boldsymbol{\beta}_i$. When no conjugate prior can be found, or when there is no convenient way of sampling from the target (posterior) distribution, the random-walk Metropolis-Hastings (MH) algorithm is a very versatile and flexible tool in Bayesian methods for sampling from these (posterior) distributions (Roberts and Casella 2005). We used a MH update step to sample from the posterior parameter space of $\boldsymbol{\beta}_i$ as there is no conjugate

prior for (4.34).

Mechanically, the MH algorithm has to have a proposal distribution and often the proposal distribution is a normal distribution. To get good performance, we want the proposal for a single $\beta_i^{(t)}$ to be neither too close nor too far from the current sample $\beta_i^{(t-1)}$, and therefore we tune our proposal density such that acceptance rates of our proposals are between 44% and 22% for single and multidimensional parameter sets respectively (Gelman et al. 2004 pg. 306).

When there is more than one component to the vector of unknowns in the MH algorithm, the component vector can be divided into sub vectors that are updated sequentially, each using its own MH proposal step. To improve convergence, sampling can be implemented by *blocking*. The component vector is divided into sub-vectors with correlated elements, called blocks, and instead each of these blocks is updated in a single MH step. Therefore, rather than using a set of uncorrelated univariate MH update steps for each unknown, it is recommended to use a multivariate normal ($\mathcal{MVN}$) proposal distribution and update the parameter vector $\boldsymbol{\beta}_i$ for each unit in blocks.

The most common proposal density is $\sim \mathcal{MVN}(\mathbf{0}, \boldsymbol{H}^{-1})$, where the tricky part is choosing $\boldsymbol{H}^{-1}$ to get the desired acceptance rates (Ntzoufras 2009). We made the diagonal and off-diagonal elements of the covariance $\mathcal{MVN}$ proposal distribution similar to the target posterior to ensure a more efficient algorithm that optimizes the acceptance rates of the proposal update values. For each fur seal, we solve for the vector of first, and matrix of second derivatives of the log likelihood function $\ell(\boldsymbol{\beta_i}) = \log(P(\boldsymbol{y}_i|\cdot))$ with respect to $l_1^{th}(l_2^{th})$ regression parameter(s) vector $\boldsymbol{\beta}_i$

$$\begin{aligned}
\frac{\partial \ell^2(\boldsymbol{\beta_i})}{\partial \beta_{il_1}^{(k)} \partial \beta_{il_2}^{(k)}} &= -\sum_j^{m_i} \left\{ X_{ijl_1} X_{ijl_2} \left(\frac{exp(\sum_{m=1}^p X_{ijm}\beta_{im}^{(k)})(1 + exp(\sum_{m=1}^p X_{ijm}\beta_{im}^{(-k)}))}{\left[1 + exp(\sum_{m=1}^p X_{ijm}\beta_{im}^{(k)}) + exp(\sum_{m=1}^p X_{ijm}\beta_{im}^{(-k)})\right]^2} \right) \right\} \\
&= -\sum_j^{m_i} \{X_{ijl_1} X_{ijl_2} h(\boldsymbol{\beta}_i)\}
\end{aligned}$$

Note here $(-k)$ is the set of non-baseline response categories with k omitted

The matrix of 2^{nd} derivatives of the multinomial log likelihood model is the Hessian matrix, and the negative, inverse matrix of the observed Fisher information matrix. The proposal distribution we consider for the MH step takes the diagonal elements $h(\boldsymbol{\beta}_{i\cdot})$ of the Hessian and the latent covariate matrix $\boldsymbol{X}_i$ at the previous time step, and combines this

with the prior variance of $\boldsymbol{\beta}_i$ (i.e. $\boldsymbol{\Sigma}_\beta$), and assuming a $\mathcal{MVN}$ prior to form matrix $\boldsymbol{H}$, the working matrix of the covariance function of the proposal distribution is

$$\boldsymbol{H}(\boldsymbol{\beta}_i^{(t)}) = \boldsymbol{X}_i h(\boldsymbol{\beta}_i^{(t-1)}) \boldsymbol{X}_i + \boldsymbol{\Sigma}_\beta^{-1}.$$

In our hierarchical, error-in-covariate framework, the covariance of our proposal distribution is then a function of (i) the information matrix of parameter vector $\boldsymbol{\beta}_{\boldsymbol{i}}^{(t-1)}$ at the last MH step, (ii) the latent lower-level covariate matrix $\boldsymbol{X}_i$, and (iii) a sample from the posterior distribution for the hyperprior of the covariance matrix, $P(\boldsymbol{\Sigma}_\beta)$.

A random walk MH update step is used where the candidate vector of coefficients is obtained by perturbing the existing vector of coefficients $\boldsymbol{\beta}_i^{(t-1)}$ by a draw from a multivariate normal distribution with means equal to the vector $\mathbf{0}$, and variance-covariance proportional to the inverse of $\boldsymbol{H}(\boldsymbol{\beta}_i^{(t)})$ which includes a term that corresponds to the covariance of the current MH draw (this is what makes it a "random walk" algorithm). Thus, our random walk MH proposal density for the i^{th} fur seal is

$$\boldsymbol{\beta}_i^{(t)} \sim \mathcal{MVN}\left(\boldsymbol{\beta}_{\boldsymbol{i}}^{(t-1)}, c^2\left[\boldsymbol{H}(\boldsymbol{\beta}_i^{(t)})\right]^{-1}\right),$$

where c^2 is the tuning parameter and set to be $2.93/\sqrt{m}$, and m is the dimension of the parameter vector (Roberts and Rosenthal 2001). We accept or reject the proposed update for $\boldsymbol{\beta}_{\boldsymbol{i}}^{(t)}$ with probability equal to α according to the Metropolis ratio between the last accepted values in the Markov chain $\boldsymbol{\beta}_i^{(t-1)}, \boldsymbol{B}^{(t-1)}, \boldsymbol{\Sigma}_\beta^{(t-1)}$ and the proposal distribution $\boldsymbol{\beta}_i^{(t)}, \boldsymbol{B}^{(t)}, \boldsymbol{\Sigma}_\beta^{(t)}$, with a correction in the MH ratio for the asymmetry of our proposal distribution.

$$\begin{aligned}
\alpha &= \min(0, r) \\
r &= log\left(\frac{P(\boldsymbol{y}_i|\boldsymbol{X}_i^{(t)}, \boldsymbol{Z}_i, \boldsymbol{\beta}_i^{(t)})\ P\left(\boldsymbol{\beta}_i^{(t)}|\boldsymbol{B}^{(t)}, \boldsymbol{\Sigma}_\beta^{(t)}\right)}{P(\boldsymbol{y}_i|\boldsymbol{X}_i^{(t-1)}, \boldsymbol{Z}_i, \boldsymbol{\beta}_i^{(t-1)})\ P\left(\boldsymbol{\beta}_i^{(t-1)}|\boldsymbol{B}^{(t-1)}, \boldsymbol{\Sigma}_\beta^{(t-1)}\right)}\right) + \\
& \quad log\left(\frac{f\left(\boldsymbol{\beta}_i^{(t-1)}|\boldsymbol{\beta}_i^{(t)}, \boldsymbol{H}(\boldsymbol{\beta}_i^{(t)})\right)}{f\left(\boldsymbol{\beta}_i^{(t)}|\boldsymbol{\beta}_i^{(t-1)}, \boldsymbol{H}(\boldsymbol{\beta}_i^{(t-1)})\right)}\right)
\end{aligned} \tag{4.35}$$

The last term in the ratio, r, of Equation (4.35) corrects for the asymmetry of the proposal density in which the numerator describes $P(\boldsymbol{\beta}_i^{(t-1)} \rightarrow \boldsymbol{\beta}_i^{(t)})$, and the denominator describes

$P(\boldsymbol{\beta}_i^{(t)} \rightarrow \boldsymbol{\beta}_i^{(t-1)})$, and is included because the covariance matrix of the step $\boldsymbol{\beta}_i^{(t-1)} \rightarrow \boldsymbol{\beta}_i^{(t)}$ is not equal to the covariance matrix used in the inverse step $\boldsymbol{\beta}_i^{(t)} \rightarrow \boldsymbol{\beta}_i^{(t-1)}$ (Ntzoufras 2009). This step-by-step adjustment of the proposal distribution means the algorithm is adaptive, as it includes the MCMC samples for the other simulated random variables in the model, and thus optimizes the acceptance ratio, r, at each step as the shape (covariance) of the proposal distribution.

4.5.3 Lower-level model for covariates measured with Berkson measurement error : A Random-Walk Metropolis step for $\boldsymbol{X}_i$

At the next stage of sampling the components of the MCMC sampling scheme for the joint posterior of Equation (4.26), we construct a Metropolis-Hastings step for sampling from the conditional distribution of (4.29), i.e.

$$P(\boldsymbol{X}_{ij}|\, \boldsymbol{y}_{ij}, \boldsymbol{W}_{ij}, \boldsymbol{Z}_{ij}, \boldsymbol{\beta}_i, \sigma^2_{iU}) \;\propto\; exp\left\{ \boldsymbol{y}_{ij}\left(\boldsymbol{\beta}_{iZ}\boldsymbol{Z}_{ij} + \boldsymbol{\beta}_{iX}\boldsymbol{X}_{ij}\right) - log\left(1 + e^{\boldsymbol{\beta}_{iZ}\boldsymbol{Z}_{ij}+\boldsymbol{\beta}_{iX}\boldsymbol{X}_{ij}}\right) + \frac{(\boldsymbol{X}_{ij}-\boldsymbol{W}_{ij})'(\boldsymbol{X}_{ij}-\boldsymbol{W}_{ij})}{2\boldsymbol{\sigma}^2_{iU}} \right\}.$$

As with the MH sampling step for the posterior $P(\boldsymbol{\beta}_i|\cdot)$, we update $\boldsymbol{X}_{i1}, \ldots, \boldsymbol{X}_{il}, \ldots, \boldsymbol{X}_{im_i}$ as a block and use a $\mathcal{MVN}$ proposal distribution. The covariance matrix of the $\mathcal{MVN}$ proposal distribution is made similar to the target posterior to optimize the acceptance rate of the proposal update values as before. We determine the Hessian matrix by finding the matrix of second derivatives of the log likelihood function $\ell(\boldsymbol{\beta}_i\boldsymbol{X}_i) = log(P(\boldsymbol{y}_i|\boldsymbol{X}_i^{(t)}, \boldsymbol{Z}_i, \boldsymbol{\beta}_i^{(t)})))$ with respect to the l^{th} measure of covariate $\boldsymbol{X}_i$, i.e.

$$
\begin{aligned}
\frac{\partial \ell^2(\boldsymbol{X}_i\boldsymbol{\beta}_i)}{\partial X_{il}^2} \;=\; & -\left[\frac{1}{1+e^{X_i\beta_i^{(k)}}+e^{X_i\beta_i^{(-k)}}}\right]^2 \times \\
& \Bigg[\left\{\frac{\partial \ell(\boldsymbol{X}_i\boldsymbol{\beta}_i^{(k)})}{\partial X_{il}}\right\}^2\left(e^{X_i\beta_i^{(k)}}\right)\left(1+e^{X_i\beta_i^{(k)}}+e^{X_i\beta_i^{(-k)}}\right) - \\
& \left\{\frac{\partial \ell(\boldsymbol{X}_i\boldsymbol{\beta}_i^{(k)})}{\partial X_{il}}\right\}\left(e^{X_i\beta_i^{(k)}}\right)\left(\left\{\frac{\partial \ell(\boldsymbol{X}_i\boldsymbol{\beta}_i^{(k)})}{\partial X_{il}}\right\}\left(e^{X_i\beta_i^{(k)}}\right)+\right. \\
& \left.\left\{\frac{\partial \ell(\boldsymbol{X}_i\boldsymbol{\beta}_i^{(-k)})}{\partial X_{il}}\right\}\left(e^{X_i\beta_i^{(-k)}}\right)\right)+ \\
& \left\{\frac{\partial \ell(\boldsymbol{X}_i\boldsymbol{\beta}_i^{(-k)})}{\partial X_{il}}\right\}^2\left(e^{X_i\beta_i^{(-k)}}\right)\left(1+e^{X_i\beta_i^{(k)}}+e^{X_i\beta_i^{(-k)}}\right) - \\
& \left\{\frac{\partial \ell(\boldsymbol{X}_i\boldsymbol{\beta}_i^{(-k)})}{\partial X_{il}}\right\}\left(e^{X_i\beta_i^{(-k)}}\right)\left(\left\{\frac{\partial \ell(\boldsymbol{X}_i\boldsymbol{\beta}_i^{(k)})}{\partial X_{il}}\right\}\left(e^{X_i\beta_i^{(k)}}\right)+\right. \\
& \left.\left\{\frac{\partial \ell(\boldsymbol{X}_i\boldsymbol{\beta}_i^{(-k)})}{\partial X_{il}}\right\}\left(e^{X_i\beta_i^{(-k)}}\right)\right)\Bigg] \qquad (4.36)
\end{aligned}
$$

Note here $(-k)$ is the set of non-baseline response categories with k omitted

We take the diagonal elements of the Hessian matrix for $\boldsymbol{X}_i$ of Equation (4.36), call this $h(\boldsymbol{X}_i)$, combine it with the prior variance of $\boldsymbol{X}_i$, σ^2_{iU}, and assume a $\mathcal{MVN}$ prior to form the inverse covariance matrix of the proposal distribution $\boldsymbol{H}(\boldsymbol{X}_i)$, i.e.

$$\boldsymbol{H}(\boldsymbol{X}_i) = h(\boldsymbol{X}_i) + (\boldsymbol{\sigma}^2_{iU})^{-1}.$$

This leads to the following random walk MH proposal distribution for covariate $\boldsymbol{X}_i$ measured for the i^{th} fur seal

$$\boldsymbol{X}_i^{(t)} \sim \mathcal{MVN}\left(\boldsymbol{X}_{\boldsymbol{i}}^{(t-1)}, c^2\left[h(\boldsymbol{X}_i^{(t-1)}) + (\boldsymbol{\sigma}^2_{iU})^{-1}\right]^{-1}\right).$$

4.5.4 Lower-level parameter for the variance of covariates measured with error : A Gibbs step for σ^2_{iU}

The Gibbs sampler is a special case of the Metropolis-Hastings algorithm in which parameters are sampled from their full conditional distributions, and thereby avoids the need to calculate any integrals.

By selecting the Inverse Gamma ($\mathcal{IG}$) density with hyperparameters a_U, b_U as a prior for the lower-level parameter describing measurement error in covariates, we make use of its conjugate properties for the variance of a normally distributed random variable, σ^2_{iU}. We can then derive its posterior distribution, and sample this target distribution directly using a Gibbs step within our MCMC chain. This $\mathcal{IG}$ posterior distribution for σ^2_{iU} is derived in Appendix (B), and is distributed as follows

$$P(\sigma^2_{iU}|\boldsymbol{X}_i, \boldsymbol{W}_i, a_U, b_U) = \mathcal{IG}\left(a_U + \frac{J_i}{2}, \quad b_U + \frac{1}{2}\sum_{j=1}^{J_i}(\boldsymbol{X}_{ij} - \boldsymbol{W}_{ij})^2\right).$$

4.6 Higher-order (Second-stage) model for population parameters; $\boldsymbol{B}, \boldsymbol{\Sigma}_\beta$

Let the following list describe the symbols that appear in the following section (and in Appendix C)

m = number of regression parameters

n = number of units/animals

q = number of higher level covariates in matrix $\boldsymbol{\mathcal{X}}$

$\boldsymbol{\psi}$ = all prior parameters of lower order parameters $(\boldsymbol{B}, \boldsymbol{\Sigma})$

$\boldsymbol{\phi}$ = all hyperparameters in the higher order model $(\nu_0, a_0, \boldsymbol{V}_0, \boldsymbol{B}_0)$

At the higher order level, we use the multivariate normal model as part of the prior structure for the lower-order regression parameters $\{\boldsymbol{\beta}\}$. We define the higher-order structure to be a multivariate linear regression model, relating the $q \times m$ higher-order regression coefficients $\boldsymbol{B}$ to the $n \times m$ matrix of lower order parameters $\boldsymbol{\beta}$ though the $n \times q$ design matrix $\boldsymbol{\mathcal{X}}$, i.e.

$$\boldsymbol{\beta}_{n\times m} = \boldsymbol{\mathcal{X}}_{n\times q}\boldsymbol{B}_{q\times m} + \boldsymbol{E}_\beta, \tag{4.37}$$

where each row of $\boldsymbol{E}_\beta \sim \mathcal{MVN}(\boldsymbol{0}, \boldsymbol{\Sigma}_\beta)$, $\boldsymbol{\mathcal{X}}$ is the $n \times q$ covariate matrix of q common (higher-level) independent variables, and the likelihood of (4.37) becomes the prior distribution for the lower-level parameters $\{\boldsymbol{\beta}\}$.

Note that this higher-order model is no longer conditional on the observed data, $\boldsymbol{y}$, $\boldsymbol{W}$ and $\boldsymbol{Z}$, but instead is only conditional on the lower-order regression parameters which are sufficient for the observed data (Gelman et al. 2004). This implies that once a set of m regression parameters is identified for each of the $i = 1, \ldots, n$ individual fur seals, these $\{\boldsymbol{\beta}_i\}$ serve as "data" for the inferences regarding $\boldsymbol{B}$, the set of parameters common to all individual fur seals.

In our application, we have no covariate data to describe the individual, thus the only component vector in the second-stage design matrix is an $n \times 1$ vector of ones (i.e. the intercepts), thus $q = 1$ and $\boldsymbol{\mathcal{X}}_{n\times 1} = \boldsymbol{1}$. This somewhat simplifies the full conditionals of the higher level parameters.

4.6.1 Gibbs steps for higher-order population parameters $\boldsymbol{B}$, and $\boldsymbol{\Sigma}_\beta$:

The next stage of the MCMC sampling scheme for the joint posterior of Equation (4.26), is to sample from the full conditional distribution of hyperparameters, $\boldsymbol{B}$, and $\boldsymbol{\Sigma}_\beta$. Since the first-stage priors were assumed to be of multivariate normal form, we use the multivariate normal regression model as the second stage likelihood in this hierarchical model setting, and write the likelihood as factored in Appendix (B) :

$$P(\boldsymbol{\beta}|\boldsymbol{B},\boldsymbol{\Sigma}_\beta,\mathbf{1}) \propto |\boldsymbol{\Sigma}_\beta|^{-(n-1)/2} exp\left\{-\frac{1}{2}\boldsymbol{S}\boldsymbol{\Sigma}_B^{-1}\right\} \times \tag{4.38}$$

$$|\boldsymbol{\Sigma}_\beta|^{-1/2} exp\left\{-\frac{1}{2}(\boldsymbol{B}-\bar{\boldsymbol{\beta}})'\mathbf{1}'\mathbf{1}(\boldsymbol{B}-\bar{\boldsymbol{\beta}})\boldsymbol{\Sigma}_B^{-1}\right\}, \tag{4.39}$$

where $\bar{\boldsymbol{\beta}} = (\mathbf{1}'\mathbf{1})^{-1}\mathbf{1}'\boldsymbol{\beta}$, or the m-column averages of the $\boldsymbol{\beta}$ matrix, and $\boldsymbol{S} = (\boldsymbol{\beta}_i - \mathbf{1}\bar{\boldsymbol{\beta}})'(\boldsymbol{\beta}_i - \mathbf{1}\bar{\boldsymbol{\beta}})$. This factored likelihood suggests an Inverse Wishart ($\mathcal{IW}$) kernel for the covariance matrix $\boldsymbol{\Sigma}_\beta$ (from 4.38), and a multivariate normal kernel for $\boldsymbol{B}$ (from 4.39). The prior on $\boldsymbol{B}$ is dependent on the scale parameters as $\boldsymbol{\Sigma}_\beta$ cannot be factored out of (4.39), thus the prior for $\boldsymbol{B}$ is a conditionally conjugate prior $P(\boldsymbol{B}|\boldsymbol{\Sigma}_\beta)$. The following conjugate priors with hyperparameters $\boldsymbol{\phi} = \{\nu_0, \boldsymbol{V}_0, \boldsymbol{B}_0, a_0\}$ were selected:

$$\boldsymbol{\Sigma}_\beta \sim \mathcal{IW}(\nu_0, \boldsymbol{V}_0) \text{ , and}$$
$$\boldsymbol{B}|\boldsymbol{\Sigma}_\beta \sim \mathcal{MVN}(\boldsymbol{B}_0, \frac{1}{a_0}\boldsymbol{\Sigma}_\beta).$$

As these priors are the natural conjugate densities we can derive the posteriors and sample directly using a Gibbs sampling step. The posterior distributions from which we sample using Gibbs steps are distributed as follows

$$\boldsymbol{\Sigma}_\beta \sim \mathcal{IW}\left(\nu_0 + n, \boldsymbol{V}_0 + \left\{(\boldsymbol{\beta}_i - \mathbf{1}\tilde{\boldsymbol{B}})'(\boldsymbol{\beta}_i - \mathbf{1}\tilde{\boldsymbol{B}}) + (\tilde{\boldsymbol{B}} - \boldsymbol{B}_0)' a_0 (\tilde{\boldsymbol{B}} - \boldsymbol{B}_0)\right\}\right) \text{ , and}$$

$$\boldsymbol{B}|\boldsymbol{\Sigma}_\beta \sim \mathcal{MVN}\left(\tilde{\boldsymbol{B}}, \frac{\boldsymbol{\Sigma}_\beta}{(n+a)}\right),$$

$$\text{where } \tilde{\boldsymbol{B}} = \frac{n\bar{\boldsymbol{\beta}} + a_0\boldsymbol{B}_0}{n + a_0}, \text{and } \bar{\boldsymbol{\beta}} = (\mathbf{1}'\mathbf{1})^{-1}\mathbf{1}'\boldsymbol{\beta}.$$

We refer the reader to Appendix (B) for the derivations of target densities, $\boldsymbol{\psi} = \{\boldsymbol{B}, \boldsymbol{\Sigma}_\beta\}$, as well as all lower-level model unknowns.

4.7 Steps of a MCMC hierarchical multinomial model with error in covariates algorithm

Below are a set of steps to implementing an error-in-covariates Bayesian multinomial logit regression:

1. Set up the model.
   (a) Identify covariates with measurement errors, and select the error model (e.g. Berkson)
   (b) Identify the likelihood.
   (c) Select priors for all unknowns in the model. Using too-imprecise priors in MCMC fitting of hierarchical models can cause problems of drifting of chains. Selecting well-considered and more precise priors can help to eliminate this difficulty (Dey et al. 1998).
   (d) Compute complete conditionals for all unknowns.
2. Perform MCMC, where the MCMC algorithm is built as a modular algorithm that sequentially generates samples from the marginal conditional distributions derived from the joint target (posterior) distribution of Equation (4.26)
   (a) Generate starting values for all unknowns.
   (b) Draw from $P(\boldsymbol{\Sigma}_\beta|\cdot)$ using Gibbs.
   (c) Draw from $P(\boldsymbol{B}|\boldsymbol{\Sigma}_\beta, \cdot)$ using Gibbs.
   (d) For each $i \in \{1, \ldots, n\}$ individuals, draw samples of lower-level regression parameters from $P(\boldsymbol{\beta}_{iZ}, \boldsymbol{\beta}_{iX}|\cdot)$ in blocks using random walk Metropolis-Hastings step.
   (e) For each $i \in \{1, \ldots, n\}$ individuals, draw samples of lower-level regression covariates measured with error from $P(\boldsymbol{X}_i|\cdot)$ in blocks using another random walk Metropolis-Hastings step.
   (f) For each $i \in \{1, \ldots, n\}$ individuals, draw an updated sample from the posterior distribution of the measurement error $P(\sigma^2_{iU}|\cdot)$ for the l^{th} covariate using Gibbs.

3. Repeat steps 2(b) through 2(f) until convergence of chain to a stationary distribution by monitoring trace plots (Gelman et al. 2004), and $\hat{R}$ (Gelman and Rubin 1992, Brooks and Gelman 1998). Discard samples prior to convergence as a burn-in period of sampling.

4. Continue sampling conditional densities until the posterior probability space is fully sampled. Thin chain as necessary such that MCMC samples from the posterior densities are independent.

5. Run multiple parallel chains (following steps 2 through 4) with different and overdispersed starting values (Kass et al. 1998). Monitor and summarize inferences on unknowns using the output from all MCMC chains to check convergence. The sampled values thinned from the target posterior distribution of all chains combined are used to calculate summary statistics of the marginal distributions of the unknowns, conditional on the observed data. From these posterior samples, calculate the uncertainty in each unknown using the 2.5^{th} and 97.5^{th} percentiles to approximate the 95% credible interval of each unknown in the hierarchical model. Other desired summaries (e.g., mean, median, mode, variance, interquartile range, etc.) of the posterior distribution of unknowns can also be calculated from the sampled values to further summarize posterior knowledge.

6. Assess various statistics describing the model's goodness of fit (Fidell and Tabachnick 2006). Is it a good model? Is a simpler model better? For hierarchical models, model adequacy investigations will be more illuminating when well-considered, informative prior specifications are selected (Dey et al. 1998).

4.8 Discussion

The methods described and implemented here offer a hierarchical Bayes framework for incorporating the outputs of a state space movement process model into an environmental process model. These methods link the use of relatively sophisticated modern tracking devices (GPS tags, archival tags and VHS tags) collecting high density data to a statistical toolbox that provides inference and understanding about behaviour of a population of diving marine predators. These analysis methods focus on the spatial tracking of a diving marine mammal and present a novel way to link each individual's behavioural patterns to the ocean habitat in which the animals forage. The hierarchical framework of this approach allows for the incorporation of multiple sources of uncertainty in our knowledge about the spatial field and about differences between individuals. By modelling these differences between individuals and using these models of each individual's behaviour to build a higher-level model that summarizes the individuals, we have provided the potential for understanding behaviour of a population of marine mammals.

4.8.1 Comparison with existing literature

Central to the spatial analysis of at-sea marine mammal movement is that at-sea movement is a complex three-dimensional process that is not always well simplified into lower dimensions. The set-up requires (i) the reduction of movement data to two dimensions and (ii) the linking of data collected at very different time and spatial scales.

One such simplification is Area Restricted Search (ARS) which aims to quantify behavior where predators slow down and increase their turning rate to (presumably) increase the encounter rate with prey (Kareiva and Odell 1987). This is a method based on calculating a straightness index that describes the sinuosity of the path by dividing the straight line distance between the start and end locations by the total horizontal distance travelled (Weimerskirch et al. 2007). The index is derived from two-dimensional tracking data, and behaviour is thereby inferred without referring to the third dimension (i.e. dive depth data; Thums et al. 2011). However, the literature has found evidence of ARS behaviour to be rare in practise (Thums et al. 2011), and others (e.g. Austin et al. 2006, Weimerskirch et al. 2007) found as many feeding events along relatively straight sections of track as in more tortuous sections.

Another simple proposal for constructing simple summary measure from two-dimensional

data and is based on a similar rationale is a quantity known as First Passage Time (FPT), or the time required for an animal to swim through a circle of a fixed radius (Fauchald and Tveraa 2003). The assumption in this metric is that the longer it takes to traverse the circle, the longer the residency time in that region of the ocean, and the higher relative foraging intensity (Fauchald and Tveraa 2003, Fauchald and Tveraa 2006). There are a number of problems with using this derived characteristic of behaviour, one significant reason is the confounding of non-dive behaviours like resting (characterized by slow swimming speeds) with tortuous paths of active foraging with each of these taking similar times to traverse a circle of fixed radius (McClintock et al. 2013). Another problem is that the methods are not related to any mechanistic model, and hence there is no biological justification for the selection of the radius of the circle. In addition, they fail to capture the wealth of relevant data from 3-dimensional dives. Nonetheless, these methods have several advantages. For example, they are relatively straight-forward and transparent, they require only two-dimensional data, and they can be robust against irregularities introduced when behaviours exist only for short periods. These methods have been used successfully to link animal behaviour as described by FPT to environmental covariates (e.g. Nordstrom et al. 2013, Benoit-Bird et al. 2013).

Both of these two-dimensional approaches (ARS and FPT) have no consideration for the separation of process and measurement error. The state space framework which provides for the separation of these error terms is a natural framework for movement analysis of tag data, and has been used extensively (eg. Royer et al. 2005, Jonsen et al. 2007, Patterson et al. 2009, Jonsen et al. 2012).

Several classes of models have been proposed for implementing this sort of approach. The first class of models to consider are state space switching models that switch among discrete behaviour models where each behaviour mode is decided by a static set of parameters, some of which could be environmental. These models work well in systems with few states, such as one in which only two behavioural states are described (e.g. transiting and foraging), and have been used with success on a diverse set of organisms including blue whales (Bailey et al. 2009), tuna (Royer et al. 2005), grey seals (Breed et al. 2009), leatherback turtles (Jonsen et al. 2005) and 23 other top predators in the North Pacific (Block et al. 2011). Higher numbers of states, however, cause confounding of movement dynamics with environmental parameters (Kim and Nelson 1999, Breed et al. 2012), thus limiting some understanding of the complexity of the movement process. Routinely top marine predators display a diversity

of dive types that can be linked to foraging behaviours, such as U and V shaped dives, and the methods presented here are well suited to distinguishing these types of dives. One key advantage of the methods described in this thesis, therefore, is they allow for (at least) three behaviour modes and are sophisticated enough to distinguish between many kinds of diving movements.

Some have criticized these discrete time or space frameworks (both in state space models and in ARS- and FPT-based models) as they may be entirely inappropriate if the discrete time scale is not behaviourally meaningful. In many cases, discrete-time models are formulated on a regular time scale dictated by the observation process, and therefore have no specific behavioural interpretation, or the analysis-scale is hand-picked with no numerical justification or consideration of error associated with its selection. Our framework is data-driven in terms of the location of the discrete time step; it is the length of time the fur seal is observed in a behaviour mode and is therefore behaviourally meaningful.

Recent progress has been made in continuous time models. In one example, two dimensional space is partitioned into regions and a diffusion model describes movement in the state, rather than as one-dimensional discrete-time increments. Harris and Blackwell (2013) is a recent example of this in which foraging habitats are considered to have centres of attraction and repulsion, and movement between nodes is modelled as a continous-time model described by a transition matrix of finite behavioural states with instantaneous probabilities of transitioning at each data point depending on the region (location) and the behavioural state (Harris and Blackwell 2013). Similar to Jonsen et al. (2007), the switching process is a finite-state Markov chain and the number of behaviour states is limited to small numbers. A key assumption is that attractive areas are discrete regions surrounded by undesirable space where movement properties depend on location and behaviour, and movement between desirable regions doesn't happen randomly. These methods hold promise and the idealised simulated tests seem to work, but as yet these methods have not been vetted on real movement data in real (and complex) spatial fields.

A similar approach was used by Johnson et al. (2008). They used continuous time analysis of tag data using the Ornstein-Uhlenbeck process that allows for a two-parameter system of random, and undirected motion superimposed on a directional drift to describe animal movement. They used the Ornstein-Uhlenbeck process model to describe, through velocity and position, correlated movement in a 2^{nd}-order autoregressive continuous time framework. Their analyses relied on lower density data such as that accompanying GPS

location data, and not the 2-second increments of an archival tag. They have not yet been formulated to include environmental covariates to inform location parameters.

One obvious modelling direction that would match time scales of behaviour with spatial scales of the environmental field is by using the *in situ* data collected by the tag itself. For most tags this includes some subset of light intensity, sea surface temperature, and salinity. This approach was taken by Bestley et al. (2013) who developed the methods for incorporating the movement process model directly into a hierarchical Bayesian state space modelling framework using these data. Bestley et al. (2013) were able to link simplified movement (resident vs. non-resident behaviour) to at-sea environmental covariates derived from the tags at relatively long intervals of 12 hours. This is certainly a positive direction from a modelling perspective and holds exciting prospects for understanding fine scale movement. However, the covariate data derived from a tag is very limited, and methodology for relating salinity and sea surface temperature to large scale processes at scales useful for management purposes has yet to be developed. This can present a major limitation of these methods depending on the scope and goal of the research.

The scale of spatial and temporal measures that characterise the environment can represent a significant source of uncertainty. This is true of any analysis that aims to link animal movement to the spatial environment. The amount of uncertainty in the data can depend on a number of factors. For example, spatial data can be collected from ships running transects across time and space, or can be derived from remotely sensed satellite data. Each data collection agency usually has its own way of pre-processing the data to account for limitations of the sampling method. For example, kernel smoothers have been applied to ship-board surveys (e.g. Benoit-Bird 2013), and 8-day composites are applied to account for cloud cover in some satellite-derived data (e.g. NOAA CoastWatch products). None of the models to date have considered the uncertainty in the acquired set of covariates. Yet the data uncertainty in modelling behavioural response to environmental covariates is omnipresent and unavoidable. The hierarchical model proposed in this chapter is ideal for incorporating uncertainty at all levels. As data quality improves, as well as our understanding of the underlying environmental processes, these methods are flexible enough to allow for better, more mechanistic, and less error-prone models of behaviour.

Incorporating parameters that relate environmental conditions to movement parameters

has been a long-standing goal of mechanistic movement model development. The spatial-temporal dynamics that describe an environmental field have complicated spatial and temporal structure, as well as spatio-temporal interactions (Berliner 1996). The data collection and modelling of these interactions is not straightforward, and can themselves only be approximated. This usually requires a whole suite of assumptions. The analysis methods discussed in this chapter average over some of the underlying complicated structure. Care must be taken so as not to make potentially unrealistic simplifying assumptions. By including measurement error in the covariates and averaging over a section of time and space, we have indirectly accounted for heterogeneity in the underlying environmental field over the length of a segment. These methods present a simplified hierarchical Bayesian model that fits the context for which these methods were developed (with coherent segments of behaviour long enough to remove any temporal/spatial autocorrelation). Furthermore, this framework is flexible enough to include many different ways of including spatial dependencies (see Gelfand 2012 and examples therein). For example the process model of Section (4.3) could include a spatial random effect with structured dependence (*cf*. Banerjee et al. 2004 for a full discussion). We emphasize that the hierarchical modelling method of this chapter is a general one that is adaptable to further expansions and refinements while still keeping within the same modelling framework.

4.9 Concluding Remarks

We have presented in this Chapter a Bayesian hierarchical linear model with error-in-covariates using an MCMC Metropolis-within-Gibbs Sampling algorithm. There is mounting evidence to demonstrate the potential perils of inferring animal behaviour based on horizontal trajectories alone (e.g. McClintock et al. 2013 and references therein). The framework introduced here is one of a set of emerging flexible, analytical techniques for extracting more out of the increasing wealth of information afforded by recent advances in 3-dimensional biologging technology. We show how to incorporate this understanding of movement into the construction of a population model for drawing inference about behavioural responses to changes in the at-sea environment. In the next chapter, we apply this hierarchical framework to a data set of eleven instrumented northern fur seals and analyze each animal's tag-data from her at-sea voyage through the Bering Sea. By fitting the model described in this chapter, we learn that fur seal behaviour in the Bering Sea is affected by both spatial

and temporal variables.

4.10 Literature Cited

Agresti, A. 1990. Categorical Data Analysis. John Wiley and Sons, New York. pp 405.

Austin, D., W.D. Bowen, J.I. McMillan, and S.J. Iverson. 2006. Linking movement, diving, and habitat to foraging success in a large marine predator. Ecology, 87: 3095-3108.

Bailey, H., B.R. Mate, D.M. Palacios, L. Irvine, S.J. Bograd, and D.P. Costa. 2009. Behavioural estimation of blue whale movements in the Northeast Pacific from state space model analysis of satellite tracks. Endangered Species Research, 10: 93-106.

Banerjee, S., B.P. Carlin, and A.E. Gelfand. 2004. Hierarchical Modeling and Analysis for Spatial Data. Chapman & Hall/CRC, Boca Raton.

Bekker, P.A. 1986. Comment on identification in the linear errors in variable model. Econometrica 54 (1): 215-217.

Benoit-Bird, K.J., B.C. Battaile, C.A. Nordstrom, and A.W. Trites. 2013. Foraging behaviour of northern fur seals closely matches the hierarchical patch scales of prey. Marine Ecology Progress Series, 479: 283-302.

Berkson, J. 1950. Are there two regressions? Journal of the American Statistical Association, Vol. 45, No. 250, pp. 164-180.

Berliner, L.M. 1996. Hierarchical Bayesian time-series models. Fundamental Theories of Physics, 79: 15-22.

Bestley, S., I.D. Jonsen, M.A. Hindell, C. Guinet, and J.-B. Charrassin. 2013. Integrative modelling of animal movement: *in situ* habitat and behavioural information for a migratory marine predator. Proceedings of the Royal Society, Series B., 280(1750): 1:19.

Block, B.A., I.D. Jonsen, S.J. Jorgensen, A.J. Winship, S.A. Shaffer, S.J. Bograd, E.L. Hazen, D.G. Foley, G.A. Breed, A.-L. Harrison, J.E. Ganong, A. Swithenbank, M. Castleton, H. Dewar, B.R. Mate, G.L. Shillinger, K.M. Schaefer, S.R. Benson, M.J. Weise, R.W. Henry and D.P. Costa. 2011. Tracking apex marine predator movements in a dynamic ocean. Nature, 475: 86-90.

Breed, G.A., I.D. Jonsen, R.A. Myers, W.D. Bowen, and M.L. Leonard. 2009. Sex-specific, seasonal foraging tactics of adult grey seals (Halichorerus grypus) revealed by state space analysis. Ecology. 90: 3209-3221.

Breed, G.A. D.P. Costa, I.D. Jonsen, P.W. Robinson, and J. Mills-Flemming. 2012. State-space methods for more completely capturing behavioral dynamics from animal tracks. Ecological Modelling, 235-236: 49-58.

Brooks, S.P. and Gelman, A. 1998. General methods for monitoring convergence of iterative simulations. Journal of Computational and Graphical Statistics, 7: 434-455.

Calder, C.A., M. Lavine, P. Mueller, and J.S. Clark. 2003. Incorporating multiple sources of stochasticity in populations dynamic models. Ecology, 84: 1395-1402.

Carroll, R.J., D. Rupert, L.A. Stefanski, and C.M. Crainiceanu. 2006. Measurement Error in Nonlinear Models A Modern Perspective. (2^{nd} ed.) Boca Raton, FL: Chapman & Hall/CRC. 455 pp.

Cressie, N., C.A. Calder, J.S. Clark, J.M. Ver Hoef, and C.K. Wikle. 2009. Accounting for uncertainty in ecological analysis: the strengths and limitations of hierarchical statistical modelling. Ecological Applications, 19(3): 553-570.

Dey, D.K., A.E. Gelfand, T.B. Swartz, and P.K. Vlachos. 1998. A simulation-intensive approach for checking hierarchical models. Test, 7(2): 325-346.

Fauchald, P., and T. Tveraa. 2003. Using first passage time in the analysis of area-restricted search and habitat selection. Ecology. 84: 282-288.

Fauchald, P., and T. Tveraa. 2006. Hierarchical patch dynamics and annual movement pattern. Oecologia. 149: 383-395.

Fidell, B.G., L.S. Tabachnick. 2006. Using multivariate statistics (5th ed.). Boston: Allyn & Bacon. 1008 pp.

Gelfand, A.E., and A.F.M. Smith. 1990. Sampling based approaches to calculating marginal densities. Journal of the American Statistical Association 85: 398-409.

Gelfand, A.E. 2012. Hierarchical modelling for spatial data problems. Spatial Statistics, 1: 30-39.

Geman, S., and D. Geman. 1984. Stochastic relaxation, Gibbs distributions, and the Bayesian restoration of images. IEEE Transactions on Pattern Analysis and Machine Intelligence, 6: 721-741.

Gelman, A., J.B. Carlin, H.S. Stern, and D.B. Rubin. 2004. Bayesian data analysis (2^{nd} ed.). Boca Raton, FL: CRC Press. 668 pp.

Gelman, A., and D.B. Rubin. 1992. Inference from iterative simulation using multiple sequences (with discussion). Statistical Science, Vol 7, 457-511.

Gilks, W.R., S. Richardson, and D.J. Spiegelhalter. 1995. Markov Chain Monte Carlo in Practice. Chapman & Hall/CRC Interdisciplinary Statistics, 497pp.

Gimenez, O., S.J. Bonner, R. King, R.A. Parker, S.P. Brooks, L.E. Jamieson, V. Grosbois, B.J.T. Morgan and L. Thomas. 2009. WinBUGS for Population Ecologists: Bayesian Modeling Using Markov Chain Monte Carlo Methods. *In* D.L. Thomson, E.G. Cooch, and M.J. Conroy. (*Eds.*), Modeling Demographic Processes in Marked Populations. *In Series* Environmental and Ecological Statistics. Springer US, 883-915pp.

Gustafson, P. 2004. Measurement error and misclassification in Statistics and Epidemiology: Impacts and Bayesian adjustments Chapman & Hall/CRC.

Harris, K.J., and P.G. Blackwell. 2013. Flexible continuous-time modelling for heterogeneous animal movement. Ecological Modelling, 255: 229-237.

Hastings, W. 1970. Monte Carlo sampling methods using markov chains and their applications. Biometrika, 57(1):97-109.

Johnson, D.S., J.M. London, M.-A. Lea, and J.W. Durban. 2008. Continuous-time correlated random walk model for animal telemetry data. Ecology, 89(5): 1208-1215.

Jonsen, I.D., R.A. Myers, and J.M. Flemming. 2005. Robust state space modeling of animal movement data. Ecology, 86: 2874-2880.

Jonsen ID, R.A. Myers, and M.C. James. 2007. Identifying leatherback turtle foraging behaviour from satellite-telemetry using a switching state space model. Marine Ecology Progress Series, 337: 255-264.

Jonsen, I.D., M. Basson, S. Bestley, M.V. Bravington, T.A. Patterson, M.W. Pedersen, R.Thomson, U.H. Thygesen, and S.J. Wotherspoon. 2012. State-space models for bio-loggers: A methodological road map. Deep Sea Research II (http://dx.doi.org./10.1016/j.dsr2.2012.07.008): 1-13.

Kass, R.E., B.P. Carlin, A. Gelman, and R.M. Neal. 1998. Markov Chain Monte Carlo in Practice: A Roundtable Discussion. American Statistician, 52: 93-100.

Kareiva, P., and G. Odell. 1987. Swarms of predators exhibit "preytaxis" if individual predators use area-restricted search. American Naturalist, 130: 233-270.

Kim, C., and C. Nelson. 1999. State-space models with regime switching: classical and Gibbs sampling approaches with applications. MIT Press Books.

Lambert, P.C., A.J. Sutton, P.R. Burton, K.R. Abrams, and D.R. Jones. 2005. How vague is vague? A simulation study of the impact of the use of vague prior distributions in MCMC using WinBUGS. Statistics in Medicine, 24: 24012428.

McClintock, B.T., D.J.F. Russell, J. Matthiopoulos, and R. King. 2013. Combining individual animal movement and ancillary biotelemetry data to investigate population-level activity budgets. Ecology, 94(4): 838-849.

McCullagh, P. and J.A. Nelder. 1989. Generalized Linear Models, Second Edition. Chapman & Hall/CRC Monographs on Statistics and Applied Probability.

Metropolis, N., A.W. Rosenbluth, M.N. Rosenbluth, A.H. Teller, and E. Teller. 1953. Equation of state calculations by fast computing machines. The Journal of Chemical Physics, 21(6): 10871093.

Nordstrom, C.A., B.C. Battaile, C. Cotté, and A.W. Trites. 2013. Foraging habitats of lactating northern fur seals are structured by thermocline depths and submesoscale fronts in the eastern Bering Sea. Deep-Sea Research II, http://dx.doi.org/10.1016/j.dsr2.2012.07.010

Ntzoufras, I. 2009. Bayesian modeling using WinBUGS. Hoboken, NJ: Wiley, 492pp.

Patterson, T.A., M. Basson, M.V. Bravington, and J.S. Gunn. 2009. Classifying movement behaviour in relation to environmental conditions using hidden Markov models. Journal of Animal Ecology. 78: 1113-1123.

Priestley, M.B. 2004. Spectral Analysis and Time Series. Academic Press. London. 890pp.

Richardson, S., and W.R. Gilks. 1993. A Bayesian approach to measurement error problems in epidemiology using conditional independence models. American Journal of Epidemiology, 138: 430-442.

Roberts, C.P., and G. Casella. 2005. Monte Carlo statistical methods. Second edition. Springer, New York, New York, USA.

Roberts, G.O. and J.S. Rosenthal. 2009. Examples of Adaptive MCMC. Journal of Computational and Graphical Statistics, 18(2): 349-367.

Royer, F., J.M. Fromentin and P. Gaspar. 2005. A state space model to derive bluefin tuna movement and habitat from archival tags. Oikos, 109(3): 473-484.

Royle, J. A., and R.M. Dorazio. 2008. Hierarchical modeling and inference in ecology: the analysis of data from populations, metapopulations and communities. Academic Press. 444pp.

Schennach, S.M. 2004, Estimation of nonlinear models with measurement error. Econometrica, 72 (1): 33-75.

Shumway, R.H., and D.S. Stoffer. 2006. Time Series Analysis and Its Applications: With R Examples. Third Edition. Springer Texts in Statistics, New York. 575pp.

Stephens, D.A., and P. Dellaportas. 1992. Bayesian analysis of generalized linear models with covariate measurement error. *In* J. M. Bernardo, J. O. Berger, A. P. Dawid, and A. F. M. Smith (*Eds.*), Bayesian Statistics 4 (pp. 813-820), Oxford: Oxford University Press.

Thums, M., C.J.A. Bradshaw, and M.A. Hindell. 2011. In situ measures of foraging success and prey encounter reveal marine habitat-dependent search strategies. Ecology, 92(6): 1258-1270.

Wasserman, L. 2004. All of Statistics: A Concise Course in Statistical Inference. Springer, New York, 445pp.

Weimerskirch, H., D. Pinaud, F. Pawlowski, and C.-A. Bost. 2007. Does prey capture induce area-restricted search. A fine-scale study using GPS in a marine predator, the wandering albatross. American Naturalist, 170: 734-743.

Wikle, C.K. 2003. Hierarchical models in environmental science. International Statistical Review, 71: 181-199.

Chapter 5

An application of a hierarchical bayesian model of northern fur seal behaviour linked to biological and physical fields

5.1 Introduction

The population of northern fur seals in the Pribilof Islands, Alaska has declined dramatically during the past 35 years, and continues to decline without any obvious reason yet identified (Towell et al. 2006). Arresting the decline of the species requires an understanding of the foraging strategies of the northern fur seal (Antonelis et al. 1997), particularly for those adult females whose foraging success is also linked to pup survival. The relationship between lactating northern fur seal movement and habitat is shaped by foraging success and the physiological constraints of feeding a stationary pup left on a beach at its natal rookery. Understanding the relationship between movement, behavior and environment is a critical part of understanding population processes (Bowler and Benton 2005), and has been identified as one of the highest-level priorities by the northern fur seal conservation plan (NMFS 2007).

Success of foraging is dependent on northern fur seals finding sufficient prey on their foraging trips through a dynamic ocean environment. Previous radio and satellite telemetry

studies of northern fur seals on the Pribilof Islands have shown that feeding areas are between 160-200 km offshore and as far as 370 km from St. Paul Island (Loughlin et al. 1987).

Despite the limitations of being tied to a rookery during the pupping season, the foraging routes from the Pribilof Islands appear to radiate outward with little formal pattern, but with some rookeries displaying a rookery-specific directional bias (Robson et al. 2004, Call et al. 2008). Past telemetry has shown, in some individuals, a within-season foraging route fidelity, with greater weight gains in those females that followed this strategy (Call et al. 2008).

Physical and biological features undoubtedly influence northern fur seal movements, but the identification of the extent of these influences remains limited (Ream et al. 2005, Kuhn et al. 2010). In particular, our understanding of the overlap between the commercial fishery and active foraging of northern fur seals remains poorly understood.

Analyses of satellite-linked dive depth data (TDR/SDR tags) to physical and biological features, have provided some understanding of at-sea dive behavior of fur seals, but the horizontal resolution of the data has been too coarse to link to fine-scale foraging preferences. The environment undoubtably influences northern fur seal movements, but identification of the extent of these influences is limited (Ream et al. 2005), and behavior of northern fur seals at sea remains poorly understood. Others have correlated northern dive depth to bathymetry (Antonelis 1997, Call et al. 2008) and lunar cycle (Ream et al. 2005). Foraging behaviour has been linked to thermocline depth and surface fronts (Kuhn 2011, Nordstrom et al. 2013), and denser prey patches found off-shelf (Benoit-Bird 2013b). These results represent an important advance in our understanding of at-sea ecology; however, they have been hampered until recently by an inability to reconstruct three-dimensional tracks between satellite locations and at-sea environmental features (Harcourt and Davis 1997). Therefore, examining fine-scale movement and behaviour in the context of the physical and biological environment the animals encounter should help us to better understand fur seal ecology and survival.

Stomach contents collected in the 1960s and 1970s and scat collected starting in the 1990s have consistently shown that the dominant prey in fur seal diets is the commercial groundfish, walleye pollock (Perez and Bigg 1986, Sinclair et al. 1994, Zeppelin and Ream 2006, Zeppelin and Orr 2010). Northern fur seal satellite data has so far linked at-sea locations to gross oceanographic features (e.g. Loughlin et al. 1999, Call et al. 2008), such as the continental shelf break to the west of the Pribilofs (Goebel et al. 1991, Robson et

al. 2004, Sterling and Ream 2004). Bengtson et al. (1985) used coarse telemetry methods to locate female fur seal feeding sites to assess the interaction of fur seals with commercial trawl fisheries and found an association along the continental shelf. However, 3-D track reconstruction that would couple the fine-scale spatial overlap of the commercial groundfish fishery to fine-scale foraging activity, would help us to better understand fur seal ecology and survival.

In this chapter, we reconstruct the 3-dimensional foraging tracks of a set of northern fur seals using satellite linked archival tags, and relate the changes in movement along the tracks, to changes in their at-sea environment. Our specific goals are:

1. To use archival tag technologies recording movement at scales of seconds, in conjunction with coarse location data from Argos satellite transmitting tags, to understand how lactating fur seals partition their time between different identifiable behaviours such as active feeding, exploratory diving, and non-feeding behaviours such as sleeping, resting, grooming or surface transiting.

2. To understand how these identified behaviours can be linked to their at-sea habitat and used to understand how physical and biological factors influence their behaviour.

This paper is structured as follows. A set of vertical movement data is introduced, a suitable movement model is proposed, and a general methodological approach is outlined in Section 5.2 with details on the application of the methodology. Results are given in Section 5.3. A discussion follows in Section 5.4.

5.2 Methods

5.2.1 Instruments

During the 2005 and 2006 breeding seasons, 18 lactating Northern fur seals were captured at Reef Rookery on St. Paul Island (57.18° N, 170.27° W; 5 in 2005, 13 in 2006).

Three different tag technologies were employed in which all instruments were attached mid-dorsally using methods described in Boyd and Coxall (1992). The first of these tags was a dead-reckoner tag (Driesen & Kern GmbH, Bad Bramstedt, Germany). The dead-reckoner tags were programmed to collect data at 2-second or 5-second intervals for the full length of the foraging trip. Two of the 2005 tags were set to sample at 5-second intervals,

and the remaining 2005 and all 2006 tags were programmed to sample at 2-second intervals. The dead-reckoner was a 10-channel logger with a 32 MB flash memory that recorded time, depth, speed (using a swim paddle system), temperature, light, pitch, roll, compass heading (in three-dimensions), and body orientation (belly up or belly down).

Dead-reckoner tags cannot detect ocean currents and are best used in combination with a satellite transmitter tag to intermittently recalibrate an animal's location. We attached an ARGOS (Platform Transmitter Terminal or PTT) satellite transmitter (Spot5, Wildlife Computers, Redmond, Washington, USA) to record latitude and longitude and thereby correct for errors associated with ocean drift in the dead-reckoner route calculations.

The third tag attached was a VHF radio transmitter (A2920 Glue On, Advanced Telemetry Systems, Isanti, Minn., USA), which allowed the animals to be relocated when they returned to the rookery. All females were recaptured to remove the tracking devices and retrieve the data series logged from their at-sea foraging trips.

5.2.2 Track Reconstruction

The 2- and 5-second resolution data recorded on the dead-reckoner tags and the Argos transmitters were integrated to link swimming position to the GPS location data (Mitani et al. 2003). The 3-dimensional fur seal track was calculated based on three data logging channels. We used the channels that recorded compass bearing, depth, and speed-through-the-water to calculate the direction and horizontal speed over the water surface for each female. We pre-processed the tag's speed channel to correct for cumulative errors due to bias in the speed paddle's position. We identified zero-speed periods (sleeping) using 26 minute time windows with near-zero variance, and used a linear interpolation to re-zero speed (Figure 5.1; 26 minute windows were selected to match the scale of analysis; this issue is discussed further in Section (5.2.3)). Each foraging track was then corrected for errors associated with ocean drift by matching times of the two tags, and forcing the dead-reckoner track to lie between the (assumed to be correct) Argos locations. This was done by translating the dead-reckoner track into polar coordinates, rotating the angle of movement, and rescaling the radial coordinate to match the direction and great circle distance between consecutive Argos locations (Figure 5.2). This resulted in a 2- or 5-second resolution track reconstruction that infills between the Argos satellite locations giving a high resolution time

and space linkage for each female over her entire at-sea foraging trip. We trimmed each female's data record to include only those data in which she was at sea. The trimmed tracks ranged from 90,000 to 480,000 3-dimensional data records per animal (Figure 5.3, Table 5.1).

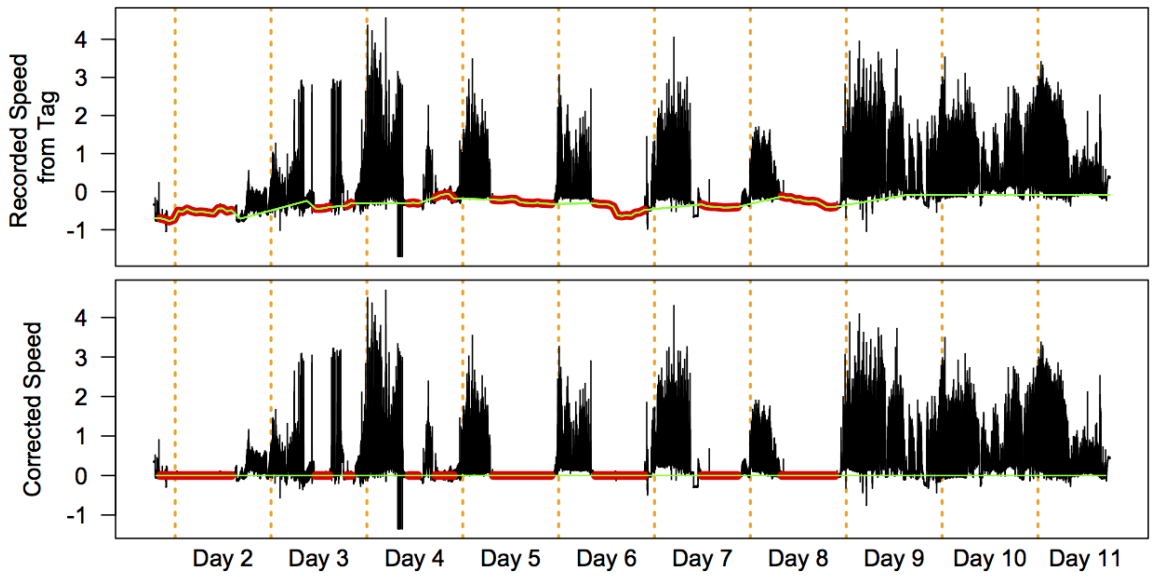

Figure 5.1: The entire raw data record for 1 female northern fur seal from the dead-reckoner's *speed-through-the-water* channel, before and after the data series is corrected for drift from zero speed. Speed is measured as speed through the water, and not over the ocean bottom, and is a tentative, approximate measure until an updated, satellite-based location can be obtained along the track. Changes in day (at midnight) are represented in this figure by the dashed vertical lines. Length of data series depicted here is 482,400 elements, and we refer to this track in the text as "Track 3".

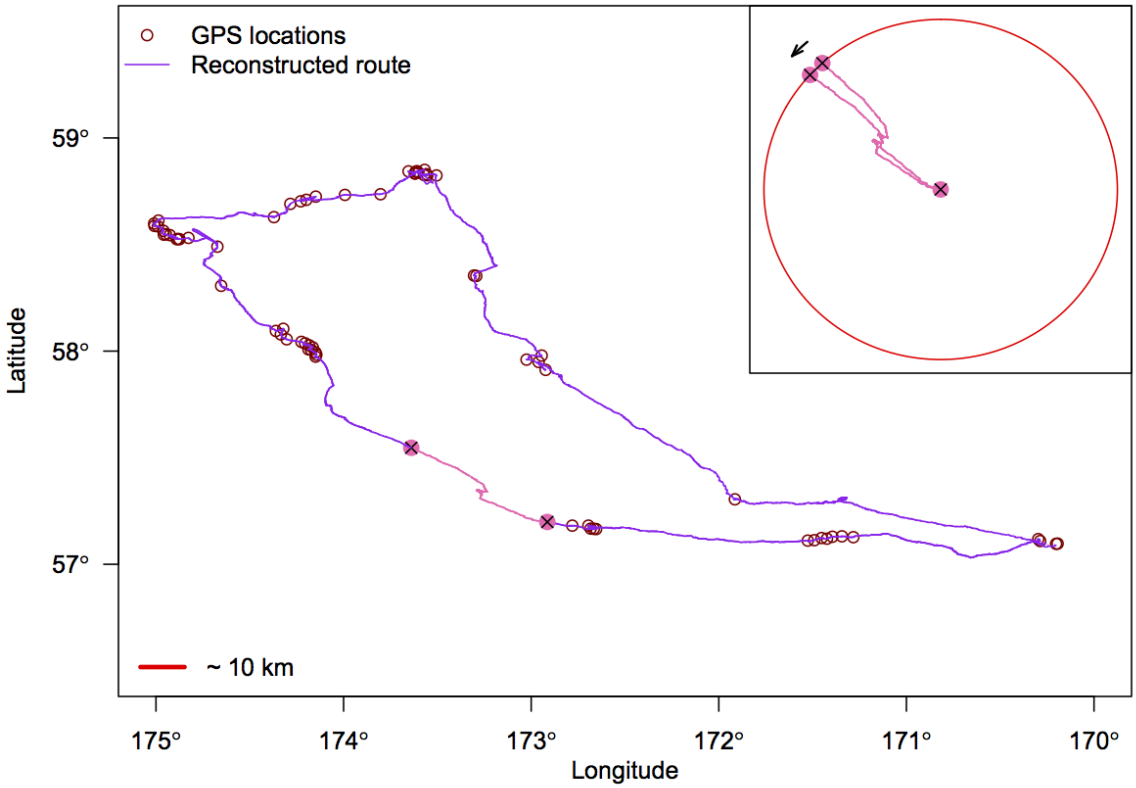

Figure 5.2: A two-dimensional reconstruction of the at-sea track called "Track 3". Constructing the best estimate of the actually track (shown as a purple line) requires connecting the high resolution archival location with the sparsely located Argos GPS locations (shown as open circles). Inset figure represents an example of the rotation step required in processing the raw dead-reckoner location data between 2 ARGOS GPS locations, thus correcting for oceanic drift from currents, and other accumulated error in the dead-reckoning between GPS fixes. The section of track in the inset figure is shown in the larger figure as a *red* portion along the out-going track. Reef Rookery on St. Paul Island is the right-hand-most location in the figure.

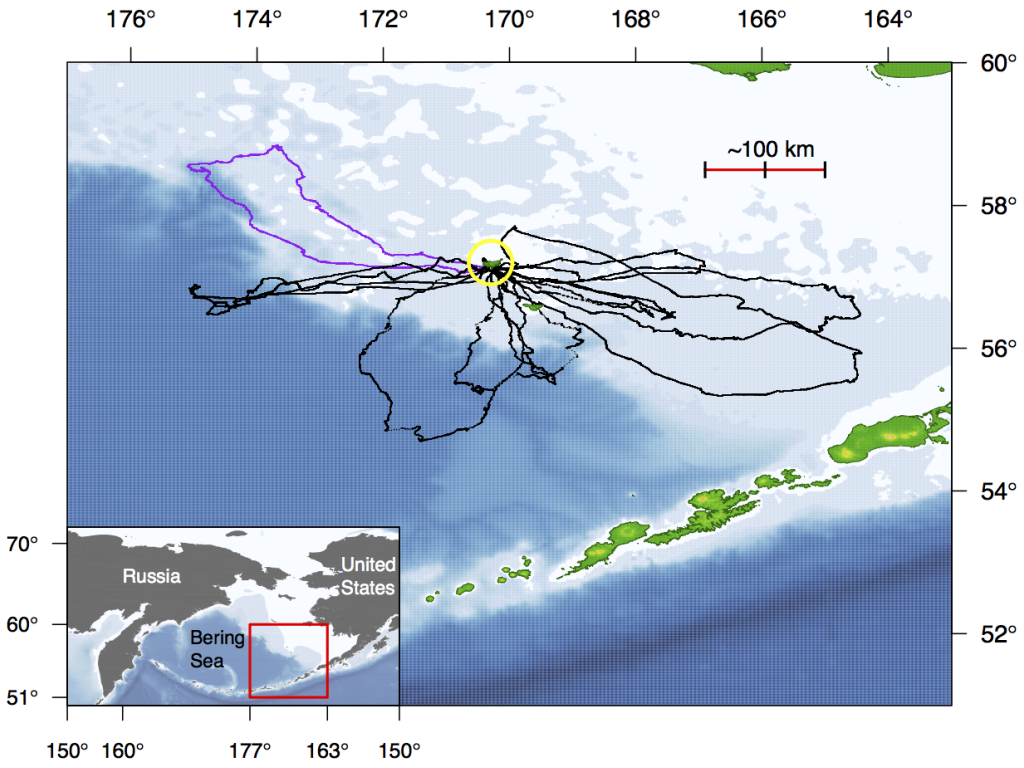

Figure 5.3: Eleven at-sea foraging tracks of lactating female northern fur seals tagged at Reef Rookery on St. Paul Island, in the Pribilof Islands in the Bering Sea, Alaska, USA. St. Paul Island is circled in yellow. The seal represented in Figure 5.2 is coloured in purple for reference. Darker shades of blue indicate deeper ocean bathymetry, in particular the transition from light blue to dark found west of St. Paul Island indicates the location of the continental shelf break (also refer to Figure 1.2 in the introduction).

5.2.3 Behavioural Response Variable

Once fine-scale locational data was resolved, we overlaid the AR(2) movement model solutions (a_1, a_2) from Section (3.1.4). We then used the results of Section (4.1.3) to translate the movement parameters to behaviour classifications based on the numerical behaviour of the AR(2) solutions. To compensate for errors in parameter estimates from unresolved tag measurement errors, and the stochastic nature of the state space solutions for similar

behaviours, we applied a kernel smoother with a locally-optimized smoothing penalty. We found an adaptable bandwidth that was locally-optimized, rather than globally optimized was best at resolving fine-scale errors. This meant the smoothing penalty was relaxed when the variance of the solutions was negligible (Herrmann 2003, R library "*lokern*").

The solution space for (a_1, a_2) identifies two distinct regions that characterize unique numerical properties (Figure 4.1). We interpret those solutions that fall in the yellow region as repetitive diving reminiscent of an oscillating pendulum, and call this behaviour mode "*Active Diving*". Parameter solutions that fall in the red area represent more intermittent diving characterized by a correlated random walk. We call the behaviour mode, "*Exploratory Diving*". These behaviour modes are shown in Figure (5.4) and correspond to the section of the time series labelled "A" and coloured *yellow*, and the section labelled "B" and coloured *red*.

Vertical velocity is small for both sleeping or resting animals and for horizontal surface movement such as transiting, and hence we could not distinguish between these two modes in our movement analysis. For purposes of the behavioural analyses, transiting, resting, and sleeping were considered the same non-diving behavioural state, as these behaviours are characterized by a lack of engagement in the immediate environment. We used the off-line estimate of system variance for each window for vertical speed (described in Appendix A) to identify non-diving states (eg. Figure 3.4). We called this behaviour mode, "*Non-Diving*".

For illustration, Figure (5.5) demonstrates the movement parameters and corresponding behaviour for Track 3 for a single day (August 18^{th}, 2006). Figure (5.6) shows how the fine-scale 3-D movement for the entire track is translated into the three behaviour modes.

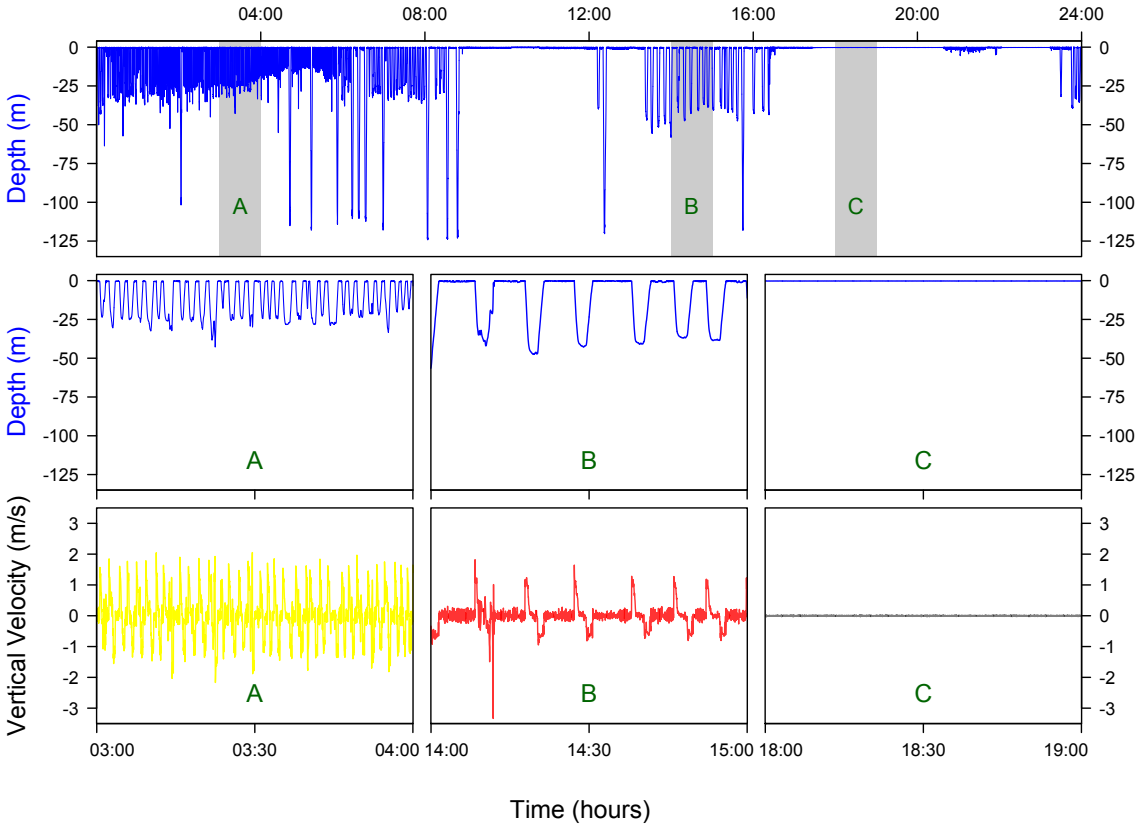

Figure 5.4: Observed time series taken from an archival tag for "Track 3" showing depth and derived vertical velocity of dives over a 24 hour period on August 18$^{\text{th}}$, 2006. Upper and middle panels show depth (in blue). The three lower figures shown in yellow, red and black, represent the vertical speed variable derived from the depth channel. These lower panels highlights three 1-hour periods stretched out accordion-style to better see the features of this high density data series and show the three northern fur seal behaviours identified in our models. Region A corresponds to a section of "'*active diving*"; Region B corresponds to a section of "*exploratory diving*"; Region C corresponds to an area of "non-diving". This data section is linked to Figures (3.1), (3.2), (3.4) and (3.3) in Chapter (3).

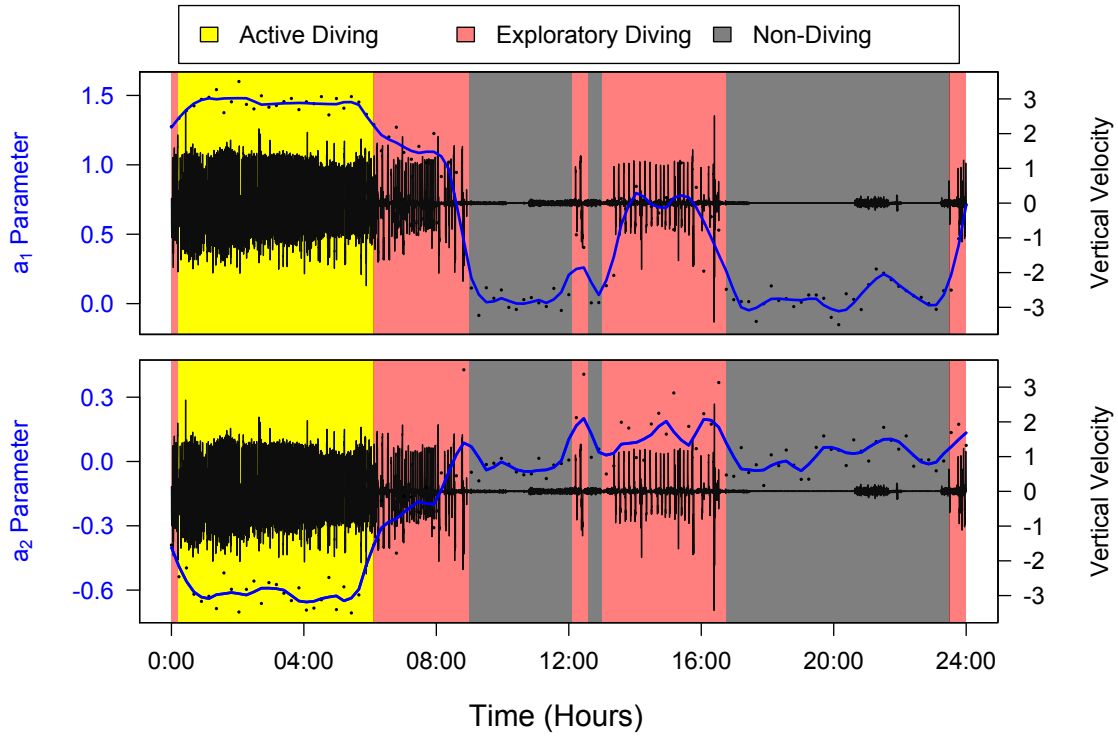

Figure 5.5: Time series for Track 3 of the smoothed movement parameters (blue line) overlaid with the original data series of vertical velocity (black lines) for a single day. These coloured blocks correspond to regions in the parameter plane for the solution space of a set of AR(2) difference equations, and regions of minimal system noise variance. The block coloured *yellow* corresponds to a region of "'*active diving*", the block coloured *pink* corresponds to an area of "*exploratory diving*", the block coloured *grey* corresponds to an area of "non-diving". Recall Figure (4.1) depicting the phase plane. See also Figure 3.4 for the associated plots of the series' system noise and observation error variances, and Figure (5.4) for the original data series for this same day, August 18^{th}, 2006.

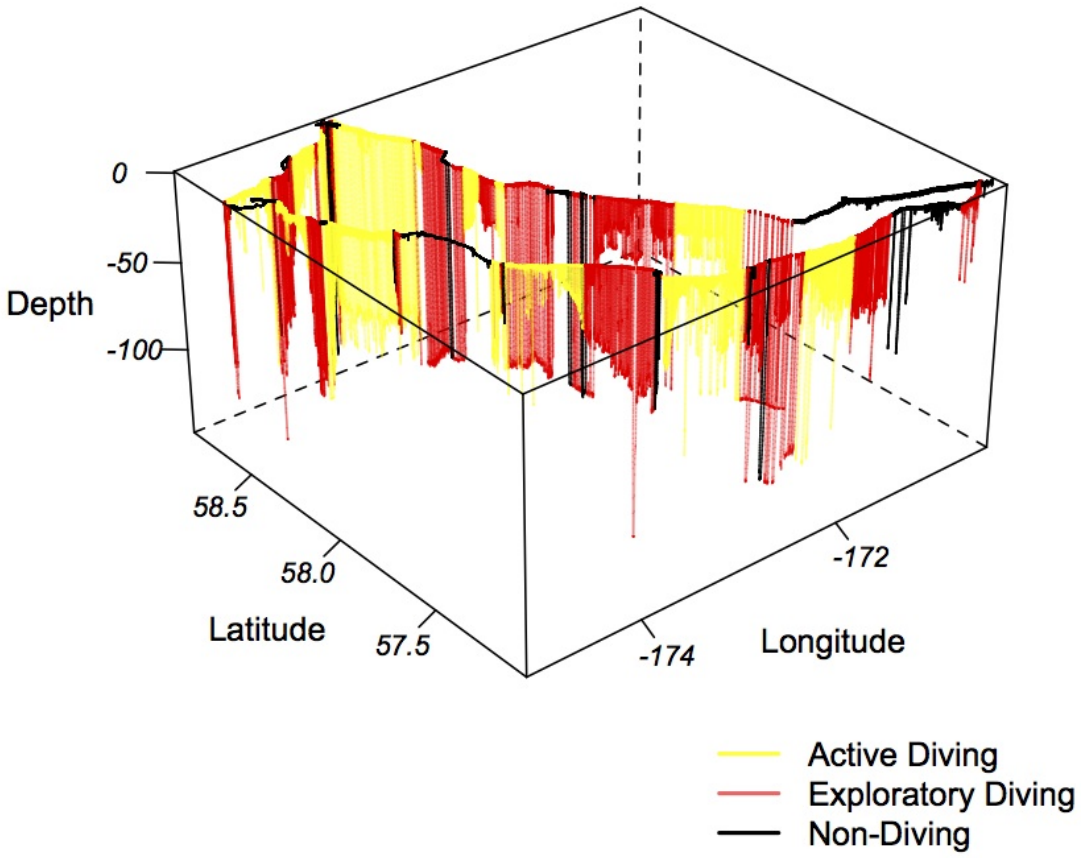

Figure 5.6: Three-dimensional reconstruction of a single northern fur seal's 11.2 day at-sea foraging track with assigned behaviour overlaid in space and time. This track corresponds to the same track (Track 3) pictured in Figures 5.1 – 5.5. Reef Rookery (start/end point) is located at the far right of this figure, the fur seal swims closer to the viewer on the way out, and returns to the rookery farther from the viewer, by way of a long transiting section.

5.2.4 Collating Spatially, Temporally Linked External Data

Time of Day

Time for each behavioural state was taken once from the dead-reckoner tag at the start of each behavioural period. As northern fur seals have strong circadian patterns in behavior (Ream et al. 2005), we transformed the time of day into a circular variable to capture this periodic rhythm, and to ensure that the interpretation of time at 24:00 is equal to that at 00:00. To do this, we modelled the dependence on time of day as a sine wave with a 24 hour period. Using a standard trigonometric identity, such a wave can be written as

$$\begin{aligned} A\ sin\left(\frac{2\pi t}{24}+\phi\right) &= A\ sin(\phi)\ cos\left(\frac{2\pi t}{24}\right)+A\ cos(\phi)\ sin\left(\frac{2\pi t}{24}\right) \\ &= \alpha_1\ cos\left(\frac{2\pi t}{24}\right)+\alpha_2\ sin\left(\frac{2\pi t}{24}\right) \quad , \end{aligned} \tag{5.1}$$

where A is the amplitude of the wave, ϕ is the phase shift, $\alpha_1 = A\ sin(\phi)$ and $\alpha_2 = A\ sin(\phi)$. Equation (5.1) provides a regression-type equation with parameters α_1 and α_2 with the sine and cosine functions serving as the explanatory variables.

Commercial Groundfish Catch and Walleye Pollock

A major goal of the study was to ascertain the extent (if any) to which fur seal behaviour was associated with fish abundance, and in particular, walleye pollock density. As a proxy for fish abundance we used the US Department of Commerce domestic observer data of the Alaska groundfish fishery for 2004-2008 (NMFS 2012). We spatially linked the fur seal tracks to each groundfish haul to the nearest 1/60th of a degree longitude and latitude (∼1.85 km). For our analysis, we selected 2 variables of interest; the first was total catch weight for each fishing haul (including both retained and discarded species), and the second was haul weight of walleye pollock. Previous stomach and scat analyses have shown walleye pollock to be dominant in their diet (Perez and Bigg 1986, Sinclair et al. 1994, Sinclair et al. 1996, Gudmundson et al. 2006, Zeppelin and Ream 2006).

We limited the fish catch data to July 9th, the beginning date of the breeding season (Trites 1992), and November 11th in 2005, and November 18th in 2006, the median dispersal dates for pups on St. Paul Island (Lea et al. 2009). For each fur seal, we linked total haul weight and pollock haul weight along the length of their foraging track. For each change

in behaviour mode along each track, we characterized the fishery by the median value of all matching locations. For example, if 10 fishing haul locations could be matched as overlapping in location with a behaviour segment of track, we took the median of those 10 haul weights to represent the groundfish catch/walleye pollock environment. The coverage of commercial groundfish catch and pollock catch was spotty, such that very few segments of track had as many as 10 matching hauls for a single summer season. It was more common for there to be only a very few matching hauls, and sometimes none. If there was no matching commercial data, the behaviour segment was omitted.

Initially, we matched fishing effort by year to our track data but found the catch data for any 1 year was too sparse to be reliable. Therefore we included all data for the years 2004 through 2008, with the acknowledgment that although small scale commercial fishing hotspots change over years, by including multiple years we had smoothed over temporal variability, but captured persistent spatial hotspots.

Sea surface temperature

Sea surface temperature data was retrieved from the dead-reckoner tag. We filtered all data that were observed outside of the top two metres of the ocean surface. For the analysis, we took the median of these surface measurements for each of the behavior states along the length of each foraging track.

Primary Productivity

We assigned primary productivity over the foraging track by constructing a composite field of weighted averages along the fur seals track by linearly interpolating time between the 8-day NOAA CoastWatch net primary productivity product
($http://coastwatch.pfeg.noaa.gov/coastwatch/CWBrowserAK.jsp$) to match the time and location of the fur seal track. This derived measure of net primary productivity is based on satellite-collected chlorophyll-a concentration and photosynthetically available radiation (PAR) measurements, corrected for the amount of organic carbon used by planktonic organisms in respiration (Behrenfeld and Falkowski, 1997), and gridded at 1/6th of a degree ($\sim$18.5 km). This was done so that each female's track was overlayed with a slightly different productivity field that was linked to the particular time range and location of each foraging trip.

Wind Speed

Also for each 24 hour period, we extracted NCDC blended satellite wind speeds at 17:00 local time (http://www.ncdc.noaa.gov/oa/rsad/air-sea/seawinds.html). This wind speed product blends satellite wind speeds from multiple platforms such as scatterometers, and passive microwave radiometers (Zhang et al. 2006). The best nominal resolution is 1/4 degree (28km) but as winds are spatially coherent on scales of hundreds of kilometres, we estimated wind speed once per day of each track, at the midpoint of the entire track and used linear interpolation across time to estimate windspeed for each behavioural segment.

Ocean Depth

As the bathymetry of this region of the Bering Sea is dominated by on-shelf shallow waters $<$ 200 metres deep and off-shelf waters with ocean depths of 3000 metres or more, we simplified bathymetry into an on-shelf/off-shelf categorical variable. Note the maximum dive depth of a northern fur seal is $\sim$200 metres (Gentry et al. 1986). For ocean depth data, we matched longitude and latitude of the track to the nearest depth measurement from the ETOPO1 bathymetry model
($http://www.ngdc.noaa.gov/mgg/gdas/gd_designagrid.html$, Amante and Eakins 2009), which gives bathymetry output in arc minutes, $1/60^{th}$ of a degree, ($\sim$1.85 kilometers). We then calculated the median depth for each of the behavior states along the at-sea track. For example if the seal was in an active diving mode between midnight and 4am, we took the median depth of the ocean during that period of the track, and classified it as on-shelf or off-shelf. We also followed methods and depth definitions of Call et al. (2008) and described depth as a three level categorical variable: inner/middle shelf (0-100 m), outer shelf (100-200 m), and off-shelf (>200 m).

5.2.5 Environmental Model

We used a post-hoc Bayesian analysis approach to examine the relationship between behaviour mode and at-sea habitat by matching the set of environmental variables in time and space to the observed behaviour model at the horizontal position of the fur seal's track. Each behaviour mode of each female fur seal is then associated in space and time to a position in the Bering Sea that changes as the fur seal moves. Our goal was to see if the coupled biological and physical variables at each location could be used to provide insight

into northern fur seal behavior. For this, we used the hierarchical Bayesian model presented in Chapter 4.

The environmental variables considered were (1) commercial haul weight of all groundfish catch, (2) haul weight of walleye pollock, (3) primary productivity, (4) sea surface temperature, and (5) wind. Total haul weight and haul weight of walleye pollock in the haul were positively correlated ($r = 0.85$), and therefore were not entered into the same model to protect analyses from matrix singularities. All five environmental variables were assumed to have measurement errors and this is accounted for in the model we fit following the error-in-covariate methods discussed in Section (4.3.1). We also considered (6) time of day and (7) bathymetry (or on-shelf/off-shelf), which were covariates that we assumed to be measured without error.

As detailed in Chapter 4, we set up a multinomial Bayesian error-in-covariate model to describe the relationship between fur seal behaviour and a set of covariates spatially and temporally matched along their at-sea foraging track. By considering segments of track in which behaviour was consistent, and by matching each segment of track to one observation for each covariate, we have averaged over both temporal and spatial autocorrelation, and the analysis is somewhat simplified. We built a lower-level parameter model to describe uncertainty in the relationship between fur seal behaviour data and the covariates, and then built a higher level model that describes the variability in the underlying process (Cressie et al. 2009). As described in Appendix (B.1), the form of the lower-level likelihood is multinomial, and the lower-level model contains unknowns $\boldsymbol{\beta}, \boldsymbol{X}, \boldsymbol{\sigma}_U$. Here, $\boldsymbol{\beta}$ are the animal-specific regression parameters, $\boldsymbol{X}$ is a set of covariates that are indirectly measured through measures $\boldsymbol{W}$ under a Berkson error model (Berkson 1950), and $\boldsymbol{\sigma}_U$ is a vector of estimated variances for the covariate measurement error in each lower-level model.

In a Bayesian setting, sampling from the posterior distribution of all unknowns requires selecting a series of prior distributions for these unknowns.

We selected the following multivariate normal ($\mathcal{MVN}$) distribution as a joint prior of the lower-level regression parameters, $\boldsymbol{\beta}$ ($cf.$ Appendix B for justification of this and all following prior distributions):

$$\boldsymbol{\beta}_i \sim \mathcal{MVN}(\mathbf{1}\boldsymbol{B}, \boldsymbol{\Sigma}_\beta). \tag{5.2}$$

As there was uncertainty in what prior distribution to select for the set of $\boldsymbol{\beta}$ regression parameters in the lower-level model, we did not assign fixed constants for the mean and

variance of the multivariate normal prior (Equation 5.2). Instead, we acknowledge this uncertainty by use of a shared hyperprior distribution that links all the lower-level models through a higher-level multivariate normal likelihood. The following forms for the hyperprior distributions were chosen

$$\boldsymbol{\Sigma}_\beta \sim \mathcal{IW}(\nu_0, \boldsymbol{V}_0) \tag{5.3}$$

$$\boldsymbol{B}|\boldsymbol{\Sigma}_\beta \sim \mathcal{MVN}(\boldsymbol{B}_0, \frac{1}{a_0}\boldsymbol{\Sigma}_\beta). \tag{5.4}$$

Equations (5.3) and (5.4) were selected to be diffuse inverse Wishart ($\mathcal{IW}$) and multivariate normal ($\mathcal{MVN}$) priors. We chose an inverse Wishart prior distribution that had parameters $\nu_0 = m + 3$ degrees of freedom (such that $\nu_0 > m$), and scale matrix $\boldsymbol{V}_0 = \nu_0 \times \mathcal{I}_m$ where $\mathcal{I}_m$ denotes an identity matrix of dimension m; i.e. Σ_β and $\mathcal{I}$ are m by m matrices. The hyperparameters in the multivariate normal distribution were selected as follows: $\boldsymbol{B}_0 = \mathbf{0}$, a length m vector of zeros, and the conditional hyperprior, $\frac{1}{a_0}\boldsymbol{\Sigma}_\beta$, with $a_0 = 0.01$ to make the scale matrix diffuse. As priors for $\boldsymbol{B}$ and $\boldsymbol{\Sigma}_\beta$ were conjugate to the higher-level likelihood, we sampled from their posteriors using a Gibbs sampling step.

We selected the following multivariate normal $\mathcal{MVN}$ and inverse gamma ($\mathcal{IG}$) distributions for the priors associated with modelling the measurement errors in the covariates

$$\boldsymbol{X}_i \sim \mathcal{MVN}(\boldsymbol{W}_i, \sigma^2_{iU}\mathcal{I}_{\mathcal{J}_i}) \tag{5.5}$$

$$\sigma^2_{iU} \sim \mathcal{IG}(a_U, b_U) \tag{5.6}$$

The prior for the *true* but unobserved covariate matrix $\boldsymbol{X}_i$ was assumed to have a multivariate normal distribution, and sampled using a Metropolis-Hastings step. We selected a conjugate inverse gamma prior for each of the covariate errors σ^2_{iU}, and used a Gibbs sampling step. We selected independent $\mathcal{IG}(a = 3, b = 1)$ priors for σ^2_{iU}, but limited the sampling range for the prior to be within a factor of 5 standard deviations of the prior distribution mean. This only affected a small sample of fur seal females for which there was limited information in the covariate data to estimate σ^2_{iU}. We fit the same, but independent $\mathcal{IG}$ priors for each of $\sigma^2_{1U}, \ldots, \sigma^2_{nU}$ fur seals, and assumed these parameters were unrelated to (and independent of) the covariate error of covariates describing the other fur seal behaviours.

In this framework, we have fit a hierarchy of models in which a set of lower-level models describing each fur seal's behaviour are then summarized in an upper-level, population model. We have therefore assumed a hierarchical framework in which our tagged animals

are assumed to be a random sample that come from a population of northern fur seals. We conduct inference at the individual and population levels by implementing an MCMC Metropolis-within-Gibbs sampling algorithm.

Sensitivity to Prior Specification

Practical implementation of Bayesian methods requires particular care, especially when the model involves weakly identifiable parameters. The problem with weakly identifiable parameters is that the two sample MCMC paths may "trade-off" their values, leading to numerical and convergence problems. A sensitivity analysis was conducted to assess the influence of prior assumptions on inferences. In particular, we focused on the sensitivity to the hyperparameter in the hyperpriors, a_0, ν_0, and $\boldsymbol{V}_0$. We also paid close attention to the observation error in the covariates σ^2_{iU} to understand how (or if) our choice of these distributional priors and hyperpriors affected inferences about northern fur seal behaviour.

MCMC Model Set-up Logistics

In seeking convergence for each model, we ran the MCMC simulation for 1,000,000 iterations, discarding the first 10,000 iterations to remove any burn-in effects. We initialised the chains from three different starting points (using different random seeds) and found posterior summary statistics were insensitive to initial conditions. Convergence is reached when the quantiles of interest for the posterior distribution do not depend on the starting values of the Markov chains and can be visually assessed by inspection of MCMC trace plots. We also calculated the Gelman-Rubin ($\hat{R}$) as modified by Brooks and Gelman (1998; using R library "*coda*") as a quantitative measure of convergence. $\hat{R}$ compares variance between and within several Markov chains run in parallel and with different starting values, and indicates convergence when $\hat{R}$ is at some low value near 1 (Kass et al. 1998). We used the partial correlation coefficient to fix the decimation rate for the MCMC chain. This is estimated by fitting autoregressive models of successively higher-orders until the lag suggests non-significant partial correlations between chain components. The thinning ratio was relatively long to ensure that the values used to produce posterior estimates were not autocorrelated; chains were thinned to every 50^{th} iteration.

We report the mean and the median of the population parameters $\boldsymbol{B}$ as measures of central tendency, and characterize the uncertainty in the marginal posterior distribution of

each unknown using the 2.5^{th} and 97.5^{th} percentiles of the posterior samples to approximate the 95% credible interval.

5.2.6 Model Comparison

Information-Theoretic Model Selection

Model comparison and selection is central to the scientific process by allowing one to evaluate different hypotheses about potential associations between variables (Pitt and Myung 2002). The best model from an *a priori* set of models and can be described as the one that explains the maximum level of detail in the simplest possible way (Burnham and Anderson 2002). Akaike (1973) introduced the Akaike Information Criterion (AIC) as a metric for the amount of Kullback-Liebler information (Kullback and Leibler 1951) lost when a model is used to approximate reality (Johnson and Omland 2004). It is written as follows

$$\begin{aligned} AIC &= -2\,\ell(\boldsymbol{y}|\hat{\boldsymbol{\Theta}}) + 2p_{\Theta} \\ &= \mathrm{D}(\hat{\boldsymbol{\Theta}}) + 2p_{\Theta}. \end{aligned}$$

Here $\hat{\boldsymbol{\Theta}}$ represents the vector of all estimated model parameters, $\ell(\boldsymbol{y}|\hat{\boldsymbol{\Theta}})$ represents the model log-likelihood and is related to the model deviance $\mathrm{D}(\hat{\Theta})$, through $\mathrm{D}(\hat{\Theta}) = -2\ell(\mathrm{y}|\hat{\Theta})$. p_{Θ} is the number of model parameters and the addition of $2p_{\Theta}$ acts as a penalty for model complexity. Spiegelhalter et al. (2002) proposed a Bayesian measure of fit or model adequacy as an alternative to AIC, known as the deviance information criterion (DIC). This metric introduces an alternate penalty term, m_{DIC}, where m_{DIC} is defined as the *effective* number of model parameters; or $m_{DIC} = \overline{D(\boldsymbol{\Theta})} - D(\bar{\boldsymbol{\Theta}})$ where $\overline{D(\boldsymbol{\Theta})}$ is the posterior mean deviance, $D(\bar{\boldsymbol{\Theta}})$ is the deviance of posterior means, and $\bar{\boldsymbol{\Theta}} = E(\boldsymbol{\Theta}|\boldsymbol{y})$ is the posterior mean of $\boldsymbol{\Theta}$, and

$$DIC = \overline{D(\boldsymbol{\Theta})} + m_{DIC}.$$

Both AIC and DIC estimate the relative information content of the models by trading off between the fit of the data to the model and the corresponding complexity of the model. One major advantage of these information-theoretic metrics is that they can be used in comparisons between non-nested models (Burnham and Anderson 2002). Burnham and Anderson (2001) recommend these approaches for the analysis of ecological data. Models that lose the least amount of information will tend to make the best predictions of replicate datasets, and have the lowest AIC and DIC values.

Both AIC and DIC can be used in a hierarchical Bayesian context depending on the focus of the modelling. In hierarchical modelling, we cannot uniquely define a 'likelihood' or 'model complexity' without specifying the level of the hierarchy that is the focus of the modelling exercise (Gelfand and Trevisani 2002), i.e. inference at the population or at the individual level. By focusing the models on a level associated with a particular set of parameters, we thereby reduce all models to non-hierarchical structures, and the counting of parameters becomes possible (Spiegelhalter et al. 2002).

The estimated AIC is the preferred metric when the modelling inference is focused on the marginal distribution of population level parameters (Speigelhalter et al. 2002). DIC instead is concerned with short-term predictive ability and is not based on any assumption of a 'true' model (Speigelhater et al. 2003). Since we were interested in understanding fur seal behaviour in general and not the behaviour (or prediction of the behaviour) of the sampled fur seals (or unit-level models), inference around $\boldsymbol{B}$ is the focus and marginal-likelihood methods such as AIC are more appropriate.

We report the posterior mean deviance $\overline{D(\Theta)}$ as a Bayesian measure of fit or "adequacy" (Dempster 1974, Spiegelhalter et al. 2002), m_{DIC} as a diagnostic of leverage (Speigelhalter et al. 2002), and DIC for comparison with AIC. However, these information theoretic metrics do not provide insight into model adequacy. They can provide useful comparisons between candidate models, but even the preferred model may not adequately fit the data.

5.2.7 Model Adequacy: Goodness of Fit and Posterior Predictive Checks

Finding distributional summaries that adequately convey the extent to which the data fit the model can be a difficult task (Cressie et al. 2009). A common and effective tool used as a diagnostic method in Bayesian analyses is the comparison of the posterior predictive distribution of replicated data under the model with the observed data (Rubin 1984; Gelman et al., 2004 p.175). If the model accurately represents the process that generated the data, then replicated data generated under the model should look similar to the observed data (Brooks and Gelman 1998). For a multinomial response model, we calculated the probability of certain behaviours given a set of observed covariates and used this to generate a sample of the posterior predictive distribution.

Let $\mathcal{L}(\boldsymbol{y}|\boldsymbol{\Theta})$ denote the likelihood for the model, where $\boldsymbol{y}$ denotes the observed response data and $\boldsymbol{\Theta}$ denotes all the parameters in the model, including the hyperpriors. The posterior

predictive distribution of replicated data $\boldsymbol{y}_{rep}$ is then defined as:

$$f(\boldsymbol{y}_{rep}|\boldsymbol{y}) = \int \mathcal{L}(\boldsymbol{y}_{rep}|\boldsymbol{\Theta}) f(\boldsymbol{\Theta}|\boldsymbol{y}) d\boldsymbol{\Theta}, \tag{5.7}$$

which is the likelihood of the future data averaged over the posterior distribution $f(\boldsymbol{\Theta}|\boldsymbol{y})$. The distribution is termed the *predictive distribution.*

The replicated data $\boldsymbol{y}_{rep}$ reflect the expected observation assuming the model with observed data $\boldsymbol{y}$. If the model is adequate, the values of $\boldsymbol{y}$ and $\boldsymbol{y}_{rep}$ should be close. An evaluation of closeness can be carried out using some summary function of discrepancy, $d(\boldsymbol{y}, \boldsymbol{\Theta})$, with assessment of the overall goodness of fit using posterior predictive p-values (Meng 1994) given by

$$p_{pp} = P\Big(d(\boldsymbol{y}_{rep}, \boldsymbol{\Theta}) < d(\boldsymbol{y}, \boldsymbol{\Theta})|\boldsymbol{y}\Big). \tag{5.8}$$

The proportion of times the summary function of discrepancy, $d(\boldsymbol{y}_{rep}, \boldsymbol{\Theta})$, is less or equal to $d(\boldsymbol{y}, \boldsymbol{\Theta})$ is then recorded across iterations of the MCMC sampler, and used as a test of the model's capacity to reproduce the observed data.

In our application, 10,000 iterations were drawn from posterior distributions after the burn-in period and used toward the calculations of predicted behaviour probabilities. We considered $d(\boldsymbol{y}, \boldsymbol{\Theta})$ to be the deviance residuals of the multinomial model. We comment that Meng (1994) and Carlin and Louis (2000) argue that posterior predictive checks provide a measure of discrepancy between the model and the data, however others (e.g. Bayarri and Berger 2000) suggest there are problems with interpreting these as p-values, and that they are not directly informative for model comparison and inference. Here, we assume that a posterior p-value around 0.5 indicates that the distributions of the replicated and actual data are close, while a value close to zero or one indicates strong differences between them (Gelman et al. 2004). Because of these unresolved differences of opinion, we use these outcomes as only one measure of goodness of fit to help capture important features of the data, but do not allow our inference to be overly dependent on these results.

Finally, Johnson (2004) developed a goodness-of-fit statistic for Bayesian model assessment, R^B. This statistic is based on outcome bins of $\boldsymbol{y}$, in which we evaluate the bin probabilities at sampled values of $\widetilde{\Theta}$ from the posterior distribution. That is if $f(\boldsymbol{y}|\Theta)$ denotes the probability mass function of a discrete random variable $\boldsymbol{y}$ for the k^{th} bin then

$$p_k(\widetilde{\Theta}) = \frac{1}{J} \sum_{j=1}^{J} f_j(\boldsymbol{y}|\widetilde{\Theta}).$$

Using draws from the posterior distributions of the MCMC sampler, R^B is calculated as follows

$$R^B(\widetilde{\Theta}) = \sum_{k=1}^{K} \frac{\left(m_k - Jp_k(\widetilde{\Theta})\right)^2}{Jp_k(\widetilde{\Theta})}.$$

where m_k represents the number of observations observed within the k^{th} bin (or behaviour type), p_k is the modelled probability of falling into the k^{th} bin, and K, the number of bins, is 3 in this instance. This statistic is asymptotically distributed as a χ^2_{K-1} distribution, if the model is accurate. Therefore, we can use R^B and the 95^{th} percentile of the null χ^2_2 distribution to calculate the proportion of time R^B exceeds this critical value, i.e. $P(R^B > \chi^2_2)$, as a measure of overall goodness of model fit.

In summary, the use of information-theoretic metrics, such as AIC, $\overline{D(\Theta)}$, does not take into account model uncertainty (Clyde 2000). These information-type metrics provide a relative measure of how well a set of model fits the data, where as checks on the posterior predictive distribution assess the ability of the model to fit the numerical properties of whole movement paths that are not explicitly included in the model. By using both AIC and a suite of posterior predictive checks, as well as monitoring MCMC convergence measures, we have controlled for some of the structural uncertainty of modeling, and can have more confidence in our inference about covariate-behaviour relationships, their predictive accuracy, and the coverage of their credible (or predictive) intervals.

5.2.8 Simulation to investigate implications of sample size

We constructed a simulation study using the Kullback-Leibler (KL) divergence statistic to examine the effect of sample size on model inference. The KL divergence has been widely studied in statistical literature as a central index measuring qualitative similarity, or relative entropy, between two probability distributions, and is popular because it uses a likelihood ratio approach (Lee and Park 2006). We took 200 samples from the posterior distribution of the population-level models to use in generating 200 replicate datasets from each sample's posterior prediction. We then refit each of the 200 datasets using the MCMC model from which the data were generated. Posteriors of these replicate simulated datasets were gathered after burn-in, from a 10,000 long chain thinned to every 5^{th} iteration. The KL divergence for each of the original model's 12 population parameter posteriors $\boldsymbol{B}$ and the 200

simulated datasets and each of these model's parameter's posteriors were then calculated using R library "$FlexMix$" (Leisch 2004).

5.3 Results

Ten of the eighteen lactating females captured at Reef Rookery yielded complete data sets from which the foraging paths and behaviour information could be retrieved. One of these animals provided data from two complete foraging trips giving us 11 complete foraging tracks. The remaining eight animals had only partial data records with lost data due to a number of tag failures including water entering the housing, the speed paddle breaking off, fish vertebrae lodged between the speed paddle and housing, and battery failure. We present the results of the behaviour analysis of the 11 complete northern fur seal foraging tracks from Reef Rookery on St. Paul Island during the pupping seasons of 2005 and 2006.

Table (5.1) summarizes the 11 foraging trips and details the collected tag data. Foraging trips ranged in length from 5.5 days to 11.2 days, with a median of 7 days spent at sea. There were an average of 40 ARGOS fixes, and 404,400 archival data records per trip.

Table 5.1: Summary of archival (data-logger) and Argos GPS tag data collected during 11 northern fur seal at-sea foraging tracks through the Bering Sea. Fur seals were tagged at Reef Rookery on St. Paul Island during the pupping seasons of 2005, 2006. The female corresponding to Track 3 who's track (or portions thereof) is visualised in Figures 5.1, 5.2, 5.3, 5.5, and 5.6 appears in **bold** font in this table. Tracks 5.1 and 5.2 correspond to the same female that took 2 trips and the tag was not recovered after her first at-sea track. Tracks 10.1, 10.2 correspond to the same tag that was deployed twice, on two separate and independent female northern fur seals.

	Female fur seal track ID											
Variable	1	2	**3**	4	5.1	5.2	6	7	8	10.1	10.2	Median
Departure date	Jul 13 2005	Jul 15 2005	**Aug 16 2006**	Aug 14 2006	Aug 13 2006	Aug 23 2006	Aug 13 2006	Aug 29 2006	Jul 16 2005	Aug 17 2006	Aug 26 2006	
Departure time	23:31	02:15	**23:28**	18:39	14:16	15:21	14:41	07:23	01:14	15:04	11:46	
Return time	12:52	19:49	**03:52**	04:55	01:49	17:09	21:34	04:11	13:12	15:29	20:42	
Trip length (days)	5.6	6.7	**11.2**	10.4	8.5	9.1	10.3	5.9	5.5	7.0	6.4	7.0
# of Argos fixes	60	15	**74**	53	27	28	54	23	31	47	40	40
# of Argos GPS fixes per day	10.7	2.2	**6.6**	5.1	3.2	3.1	5.2	3.9	5.6	6.7	6.3	5.2
# of archival tag data records	98800	291200	**482400**	451600	404400	430400	445200	240000	94560	303200	406400	404400
Sampling rate per minute	20	30	**30**	30	30	30	30	30	20	30	30	30

Figure (5.3) depicts the foraging track reconstructions highlighting the locations of these 11 foraging tracks in the Bering Sea. The fur seals traveled an average maximum linear distance from St. Paul Island of 279 km, with a maximum distance away of 391 km. Five animals stayed on the continental shelf, one animal moved along the shelf break foraging in the canyons, and 5 animals went across the shelf break into the deep water of the central Bering Sea. The one animal for which we had 2 successive trips (Tracks 5.1, 5.2) went to a similarly located feeding area off-shelf in both foraging trips. In general, these fur seals covered a wide area of both on-shelf and deep, off-shelf waters in the Bering Sea and showed no preference for foraging in any single area.

Fur seals spent 35% of the time at sea engaged in either active (14%) or exploratory dives (21%). Fur seals spent 65% of time at the surface engaged in non-diving behaviours. The 11 foraging trips had a median of 41 segments in different behaviour states, with those animals with longer foraging trips exhibiting more behaviour changes. Figure (5.7) depicts the proportion of segments in each of the three behaviour states per track. Figure (5.8) plots the distribution of segment lengths of track. This figure shows that for all 421 behaviour segments observed across all animals, the most commonly observed segment length was < 1 hour, but this was by far the most frequent time length for active diving segments (44%), whereas exploratory diving and non-diving segments were comparatively less likely to be < 1 hour (26.5% and 15.1% respectively). The median and maximum time spent in *active diving*, *exploratory diving* and *non-diving* modes is found in Table (5.2).

Table 5.2: Observed median and maximum time spent in each of three behaviour modes

	Length of Behaviour Segment	
Behaviour mode	Median	Maximum
Active diving	1 *hrs* 18 *min*	8 *hrs* 40 *min*
Exploratory diving	1 *hrs* 44 *min*	16 *hrs* 54 *min*
Non-diving	7 *hrs* 48 *min*	36 *hrs* 11 *min*

A three-dimensional reconstruction of one at-sea foraging route (Track 3) is presented in Figure (5.6) in which behaviours are identified by changes in colour.

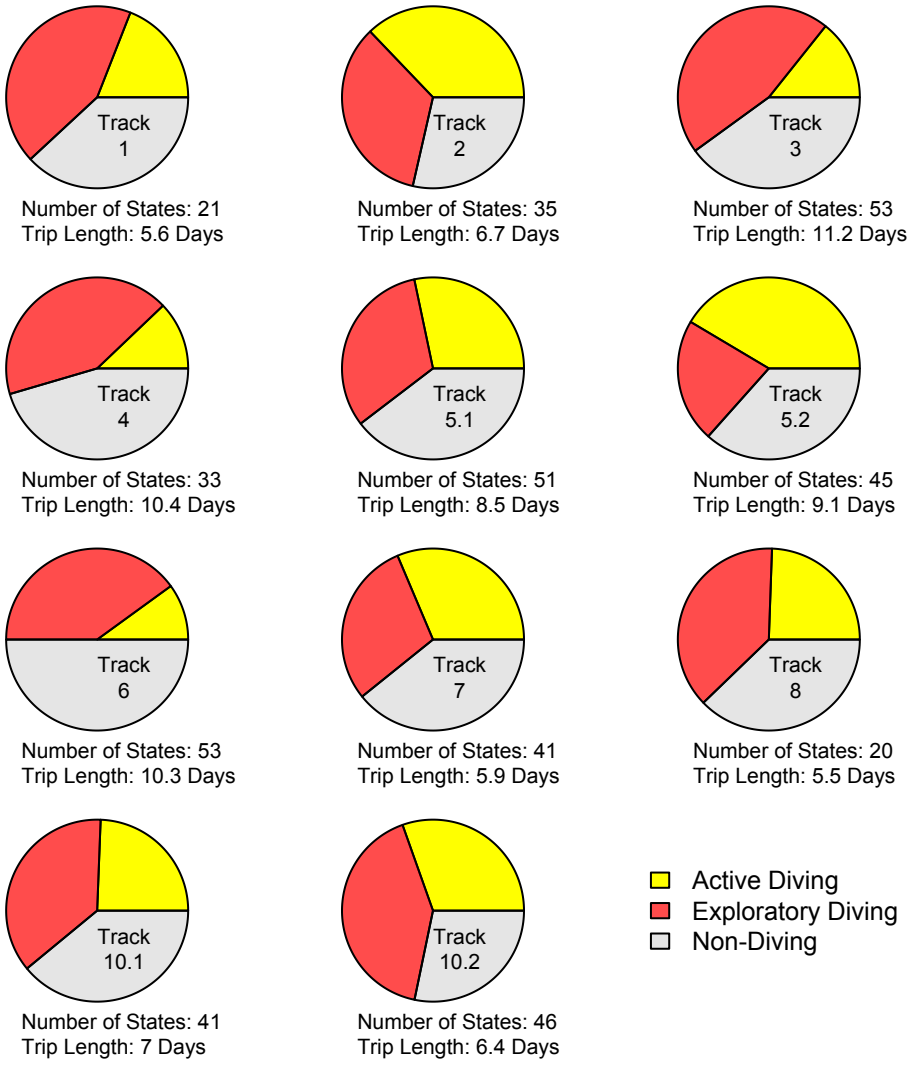

Figure 5.7: Number of behavioural states, and total trip length of 11 at-sea foraging tracks through the Bering Sea taken by female northern fur seals. Pie charts show relative number of three behaviour states: *active diving*, *exploratory diving* and *non-diving*, as well as total number of behaviour states. This figure is a visualization of the relative number of states that contributed to the model.

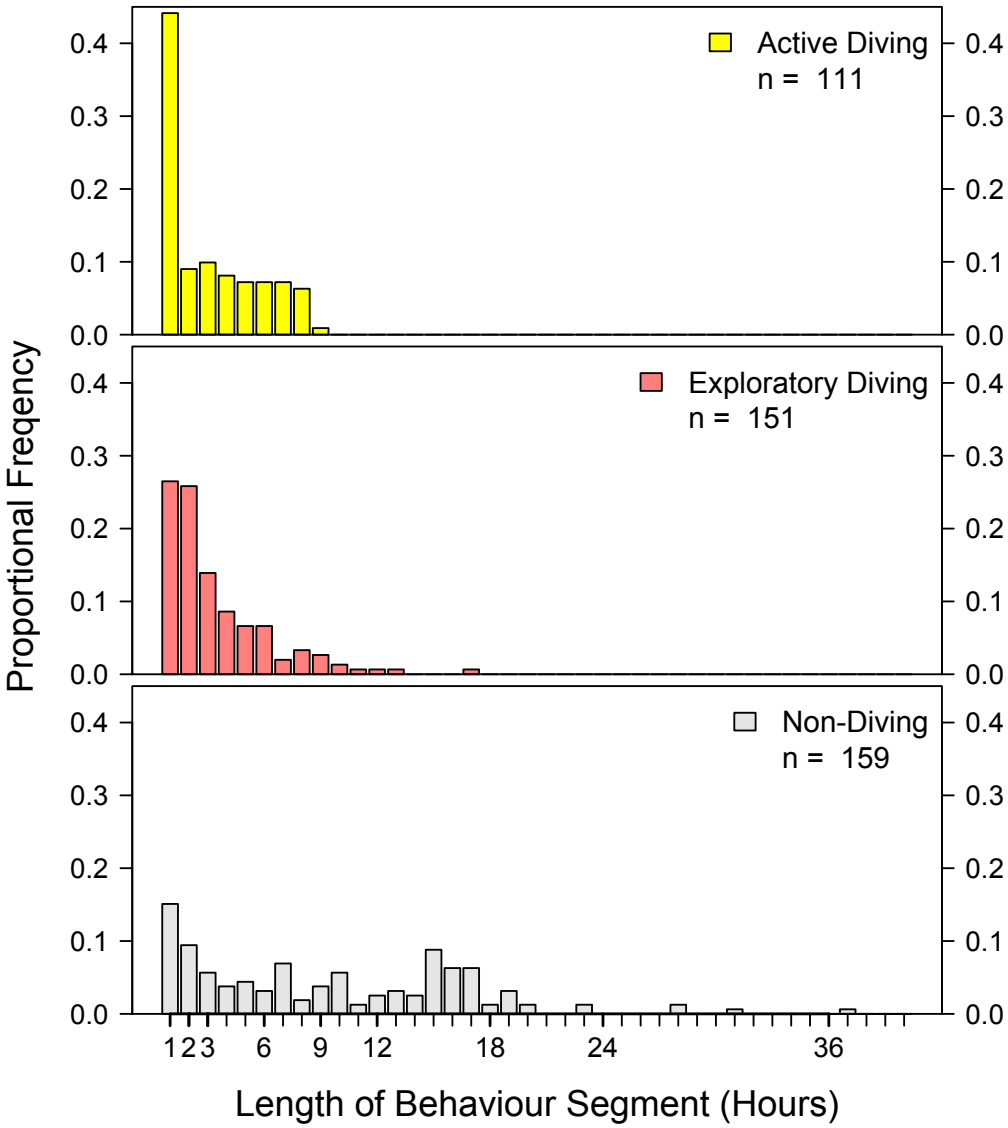

Figure 5.8: Length distribution of 421 behavioural segments measured from 11 at-sea foraging tracks of female northern fur seals. Each panel shows the distribution of one of the three behaviour states: *active diving*, *exploratory diving* and *non-diving*, where each bar represents the relative frequency of observed segments of lengths up to that hourly measure. For example, the first bar labeled "1" represents the frequency distribution of segments between 0 and 1 hour long, the second bar labeled "2" represents the frequency distribution of segments between 1 and 2 hours long, *etc.*

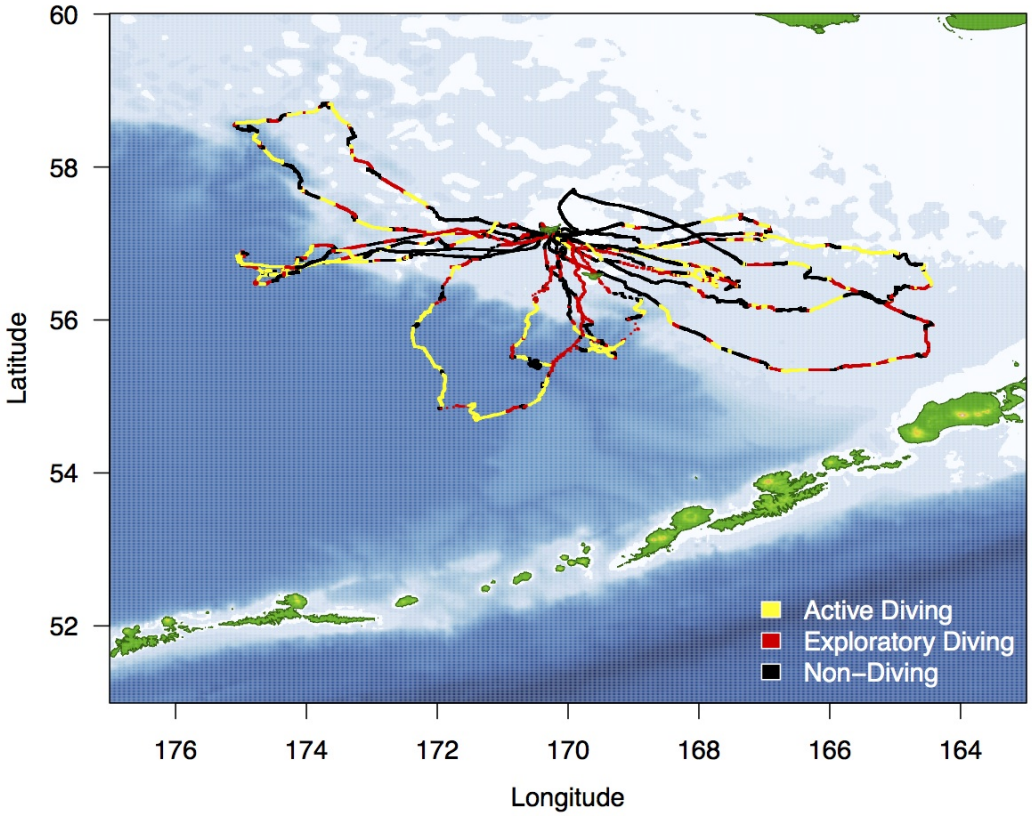

Figure 5.9: Identified behaviours of northern fur seal overlaid on eleven at-sea foraging tracks of lactating female northern fur seals from St. Paul Island, Alaska.

5.3.1 Model selection

We began our model selection procedure by fitting all main effect models with covariates described in Section (5.2.4). The strongest univariate predictor of behaviour was time of day (Table 5.3), but it was also found to be related to total (logged) haul size of all catch and (logged) haul weight of walleye pollock. We found behaviour unrelated to bathymetry, primary productivity, sea surface temperature or wind. We then fit a series of models that all included time of day and one by one inclusion of the remaining covariates. Models fit with

(logged) haul size of walleye pollock, total (logged) catch weight of the haul, and bathymetry (on/off shelf) were the best subset by the model selection criteria presented earlier. Fish hauls were related to behaviour (significant Bayesian p-value), but bathymetry was unrelated. We found that the depth categories described by Call et al. (2008) were not important and are not discussed further. The model that included time of day and primary productivity failed the posterior predictive tests (i.e. $p_{pp} \leq 0.05$, $\text{✗} \triangleq P(R^B > \chi^2_{2,0.95}) \geq 0.05$), despite having adequate higher-order parameter convergence in their respective MCMC's (i.e. $\hat{R} \approxeq 1$). We failed to get satisfactory convergence with covariates sea surface temperature and wind, and these results are omitted from Table (5.3). One additional step of model complexity was considered. We attempted a three covariate model based on the best time-interaction model, and bathymetry (on/off shelf; Time $\times$ Pollock + Bathymetry). The penalty term in both AIC and DIC penalized this model aggressively, suggesting the additional complexity was not warranted despite providing a fair goodness of fit to the data. Table (5.3) reports a subset of the results from the model fitting.

We considered a small subset of models for comparison of model fit with and without including measurement error in the covariates. In Figure (5.10), we see that the effect of including the measurement error in the model is to shrink the regression parameters towards zero, and to increase the uncertainty around the estimates of those parameters. Table (5.4) compares the model fit diagnostics, information criteria and posterior predictive checks both with and without error-in-covariates. By including the uncertainty in the measurement of haul size of pollock, the model fits significantly better by both AIC and DIC. Furthermore, ignoring the measurement error in covariates suggests the model does not fit the data well.

Our final selected model (Time $\times$ Pollock) suggests that female at-sea behavior was influenced by time of day, and (logged) haul size of walleye pollock. Time $\times$ Total haul was a similarly good model with only marginally poorer criteria based on AIC, with similar model adequacy measures ($\overline{D(\boldsymbol{\beta})}$), but a larger number of parameters estimated as implied by comparison of m_{DIC}. Coefficient magnitudes, and signs were similar for both models. Figure (5.10) and Table (5.5) depict the 95% credible intervals for each of the population parameters, and shows significance for the main effect of time of day, and for the interaction of time of day and haul size of pollock. Figure (D.2) in Appendix (D) isolates the posterior distribution that corresponds to the log odds of active diving vs. baseline non-diving states (i.e. $log(p^{(3)}_{ij}/p^{(1)}_{ij})$) for eleven lower-level parameters $\boldsymbol{\beta}_{1:3}$ and upper level parameter $B_{1:3}$ for the interaction of (sine) time of day and (logged) haul size of pollock.

Table 5.3: Model information criteria, and model fit diagnostics. *AIC* and *DIC* are relative measures of model fit, $\overline{D(\boldsymbol{\beta})}$ is a measure of model adequacy, m_{DIC} is an estimate of effective number of model parameters, and m represents the actual number of population regression parameters in the upper level model. Posterior predictive p-values $p_{pp} > 0.05$ implies there is no evidence model is predicting poorly. A posterior predictive χ^2 test shows probability of test statistic R^B exceeding $\chi^2_{2,0.95}$; ✓ denotes model as adequate, ✗ as not.

Main Effect Models (by Model Name)	*AIC*	*DIC*	$\overline{D(\boldsymbol{\beta})}$	m_{DIC}	p_{pp}	$P(R^B)$	m
Time Only	346.5	383.2	351.3	31.9	0.80	✓	6
Time + Total haul	290.1	388.8	349.9	38.9	0.55	✓	8
Time + Pollock	290.0	381.8	344.0	37.8	0.72	✓	8
Time + Bathymetry	298.7	385.3	344.5	40.8	0.84	✓	8
Time + 1° Productivity	382.0	384.2	344.0	40.2	0.04	✗	8

Time Interaction Models	*AIC*	*DIC*	$\overline{D(\boldsymbol{\beta})}$	m_{DIC}	p_{pp}	$P(R^B)$	m
Time × Total haul	285.6	373.0	322.8	48.5	0.32	✓	12
Time × Pollock	**284.3**	**368.1**	**323.2**	**44.3**	**0.72**	✓	12
Time × Bathymetry	303.6	382.6	330.8	51.8	0.62	✓	12
Time × 1° Productivity	292.0	394.8	334.7	60.0	0.00	✗	12

Three Covariate Model	*AIC*	*DIC*	$\overline{D(\boldsymbol{\beta})}$	m_{DIC}	p_{pp}	$P(R^B)$	m
Time × Pollock + Bathymetry	332.3	380.2	314.1	66.2	0.51	✓	14

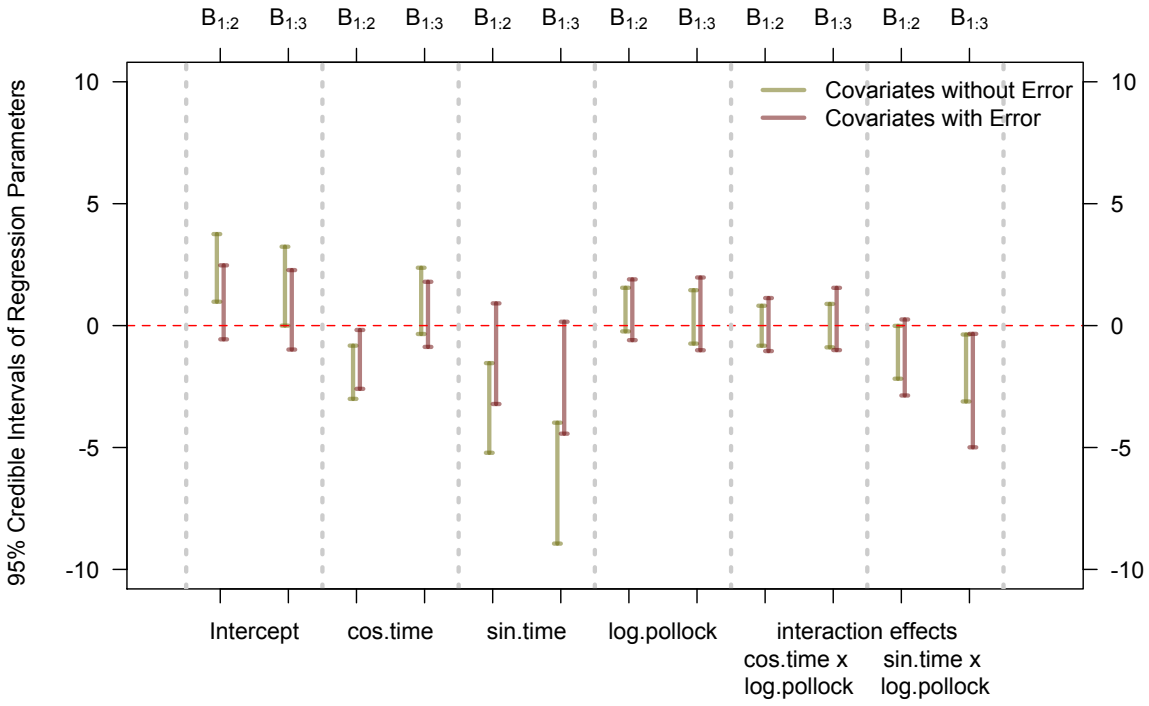

Figure 5.10: 95% credible intervals for twelve regression parameters $\boldsymbol{B}$ from the population level model that includes *time of day* and (logged) haul size of *pollock* (Time × Pollock) measured without error and with error. Each pair of $\boldsymbol{B}$ coefficients show the credible intervals of the population-level parameters linking northern fur seal behaviour to at-sea habitat. $B_{1:2}$ denotes the regression parameters corresponding to the logit response $log(p_{ij}^{(2)}/p_{ij}^{(1)})$, or the log odds of exploratory diving vs. baseline non-diving. $B_{1:3}$ denotes the regression parameters corresponding to the logit response $log(p_{ij}^{(3)}/p_{ij}^{(1)})$, or log odds of active diving vs. baseline non-diving).

Table 5.4: Comparing a selected model with and without consideration of the measurement error in covariates. This table presents model fit diagnostics, information criteria, and posterior predictive checks, for which the descriptions and definitions are comparable to those in Table (5.3).

Time Interaction Models	AIC	DIC	$\overline{D(\boldsymbol{\beta})}$	m_{DIC}	p_{pp}	$P(R^B)$	m
Time × Pollock without σ^2_U	297.0	386.0	329.4	58.8	0.461	✗	12
Time × Pollock with σ^2_U	**284.3**	**368.1**	**323.2**	**44.3**	**0.72**	✓	**12**

Table 5.5: Posterior summaries for higher level model coefficients $\boldsymbol{B}$. Similar to Figure (D.1), $B_{1:2}$ and $B_{1:3}$ denote the regression parameters corresponding to $log(p^{(2)}_{ij}/p^{(1)}_{ij})$, and $log(p^{(3)}_{ij}/p^{(1)}_{ij})$ respectively. The column labeled $\hat{R}$ is the Gelman-Rubin Bayesian measure of convergence. Parameters with significant Bayesian p-values are noted in ***bold***. $B_{1:2}$ denotes the regression parameters corresponding to the logit response $log(p^{(2)}_{ij}/p^{(1)}_{ij})$, or the log odds of exploratory diving vs. baseline non-diving. $B_{1:3}$ denotes the regression parameters corresponding to the logit response $log(p^{(3)}_{ij}/p^{(1)}_{ij})$, or log odds of active diving vs. baseline non-diving).

		Mean	St. Dev.	Median $q_{.5}$	Credible Interval $q_{.025}$	$q_{.975}$	$\hat{R}$
Intercept	$\boldsymbol{B}_{1:2}$	0.932	0.772	0.928	-0.564	2.473	1.01
	$\boldsymbol{B}_{1:3}$	0.645	0.828	0.649	-0.983	2.277	1.00
cos(time)	$\boldsymbol{B}_{1:2}$	**-1.352**	**0.613**	**-1.336**	**-2.594**	**-0.179**	**1.02**
	$\boldsymbol{B}_{1:3}$	0.422	0.680	0.413	-0.874	1.797	1.00
sin(time)	$\boldsymbol{B}_{1:2}$	-1.144	1.055	-1.146	-3.216	0.913	1.00
	$\boldsymbol{B}_{1:3}$	-2.111	1.175	-2.098	-4.434	0.159	1.04
log(pollock)	$\boldsymbol{B}_{1:2}$	0.622	0.633	0.610	-0.597	1.897	1.03
	$\boldsymbol{B}_{1:3}$	0.477	0.762	0.478	-1.011	1.974	1.03
cos(time) × log(pollock)	$\boldsymbol{B}_{1:2}$	0.064	0.549	0.068	-1.045	1.131	1.01
	$\boldsymbol{B}_{1:3}$	0.264	0.645	0.257	-1.005	1.549	1.01
sin(time) × log(pollock)	$\boldsymbol{B}_{1:2}$	-1.258	0.789	-1.239	-2.866	0.252	1.00
	$\boldsymbol{B}_{1:3}$	**-2.479**	**1.177**	**-2.419**	**-4.993**	**-0.339**	**1.05**

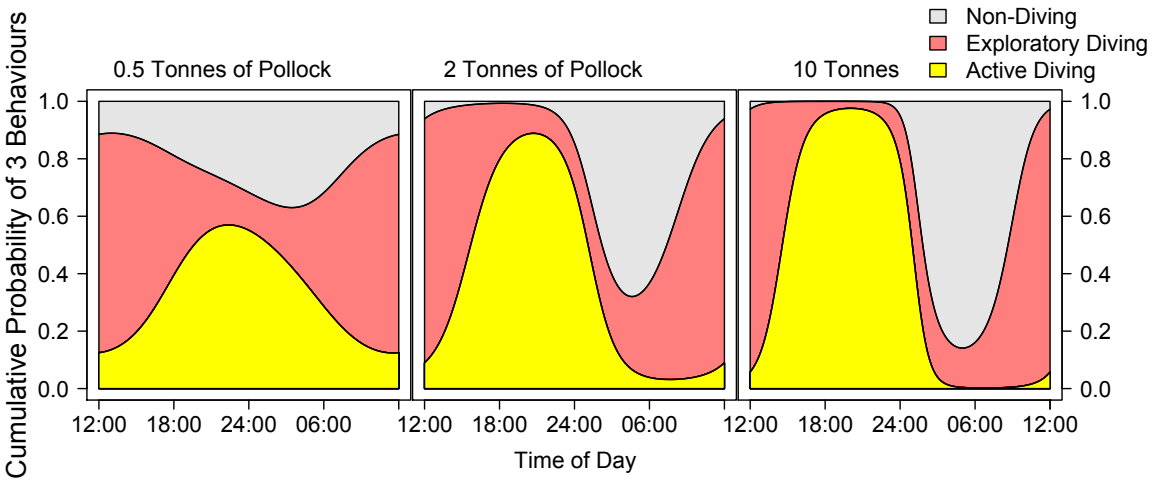

Figure 5.11: Stacked probabilities of behaviour modes *Active Diving*, *Exploratory Diving*, and *Non-Diving* in response to *time of day* and increasing commercial haul size of *walleye pollock*. The left panel of graph shows the predicted relative probabilities of behaviours in areas of small-sized walleye pollock hauls (0.5 tonnes), the middle panel shows predicted relative probabilities of behaviours in areas of medium-sized hauls (2 tonnes), and the right panel shows predicted relative probabilities of behaviour in areas of larger-sized hauls of pollock (10 tonnes). Note that the $x-$axis depicts *time of day* starting and finishing at noon to highlight the maximum amplitude of active foraging (in *yellow*) that happens at night.

Final Model

Figure (5.11) shows one daily cycle of behaviour in response to changes in time of day and commercial haul size of walleye pollock. This figure highlights the increase in relative probability of engaging in active dive behaviours in response to increasing weights of pollock haul. This is accompanied by a near-zero probability of non-dive behaviours during those phases of active diving in high catch areas. In areas of high catch, mornings ($\sim 5:00$) are characterized by higher probabilities of non-dive behaviours (surface-transitting, or resting), while low catch areas have relatively higher probabilities of dive-related behaviours at at this time of the morning. One caution in the interpretation of this figure is that our model does not predict the length of time that the northern fur seals engage in particular behaviours, but rather predicts the probability of a behaviour beginning at various times of the day (the covariate in the design matrix is *time of day* at the start of that segment of track that corresponds to that behaviour).

Equation (5.1), gives the interpretation of the regression coefficients that correspond

to time of day. From here we can calculate where in the period the peak in amplitude occurs. Note that this translation of time is a 1st-order Fourier expansion, which limits the functional form to a sinusoidal curve with a period of 24 hours. Within this restriction, the model suggests the peak in the start of exploratory dive behaviours is around noon. It suggests fur seals are most likely to begin active diving behaviours at just past 8:00 in the evening, and most likely to end diving and begin non-diving behaviours around 5:00 in the morning. Comparing results for a haul size of 10 t vs. 0.5 t suggests there is a higher probability of starting active dive behaviours in the evening, and hence a lower probability (apparently near-zero) of non-dive behaviours at that time.

5.3.2 Convergence and Sensitivity

For some models the MCMC algorithm did not converge. However, for those models that led to convergence, it was achieved within very few iterations as determined by the Gelman-Rubin $\hat{R}$ statistic. The effect of the hyperpriors was minimal on these parameters. However to get full sampling of the posterior for covariates measured with error we had to run the model for longer, and decimate the chain by keeping only every 50^{th} candidate sample. Even with this long an MCMC chain, we failed to get satisfactory convergence with a subset of covariate models (e.g. sea surface temperature and wind and these results were omitted from Table 5.3).

Within the final selected model, some of the fur seals did not have enough information in the covariates to identify both the latent unobserved covariate $\boldsymbol{X}$ and the estimate of measurement error, $\boldsymbol{\sigma}_U^2$. We experimented widely with different parameter values for the prior for $\boldsymbol{\sigma}_U^2$, and found that this prior not only affected convergence but also informed the location of posteriors for higher level parameters. The more restrictive the prior we put on the variance of the covariate error, the greater the coefficient effect and the poorer the model fit (AIC was larger). For example, a variance of zero implies a model without measurement error in the covariates, and the effect of including the error in the covariate measure is that this pulls the magnitude of the associated parameter estimate towards zero, as well as increasing the uncertainty around that estimate. In our implementation, to allow for identification in those tracks with weakly identifiable for $\boldsymbol{\sigma}_U$ and $\boldsymbol{X}$, we opted to reject any posterior samples of $\boldsymbol{\sigma}_U$ that were greater than 5 standard deviations from the expected value of the prior distribution (Figure D.4). Since the selected prior distribution was diffuse, this was not a restrictive condition except for 2 fur seal tracks, Tracks 1 and 10.2. As the priors

for each σ^2_{iU} are independent priors, this restriction on the magnitude of the measurement error will not directly affect the measurement error estimates of the other fur seal's data. In other applications, careful examination of the prior specification may impose enough constraint to allow for identification, without restricting variance components. Trace plots and pictorial evidence of convergence appear in Appendix (D).

5.3.3 Kullback-Leibler (KL) divergence and sample size

Central to the Bayesian paradigm is the notion that as the data quantity and quality increase, the posterior is less sensitive to prior assumptions (Cressie et al. 2009). The results in Figure (5.12) summarizing the KL divergence simulation described in Section (5.2.8) show there is an important gain in model performance as the number of tagged fur seals approaches 30, after which the KL divergence becomes a relatively flat function. These results therefore suggest that having data collected from $\sim$ 30 tagged northern fur seals would provide dependable posterior predictive inference, and that our sample size is too small for the data to assume such a dominant role in the assessment of how these spatial at-sea covariates affect northern fur seal behaviour (as one might suspect with only eleven tracks).

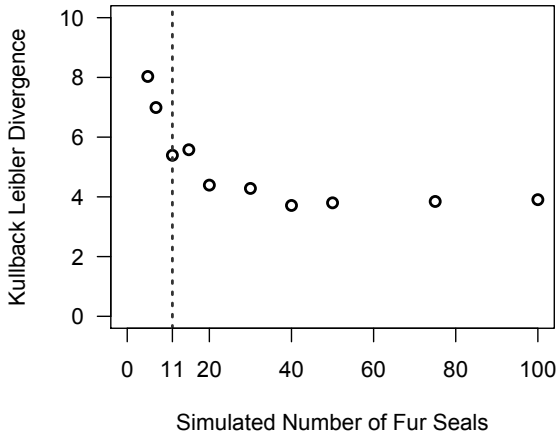

Figure 5.12: Effect of increasing sample size on Kullback-Leibler (KL) divergence for one selected population-level parameter B_{12}. A KL divergence is 0 if and only if the posterior distribution for a simulated dataset is identical to the model from which it was generated. The dashed vertical line at 11 corresponds to the sample size for this study.

5.4 Discussion

The hierarchical Bayesian framework used in this analysis has permitted us to significantly relate dive behaviours to the temporal and spatial fields through which lactating northern fur seals must forage in the eastern Bering Sea. We have establish relationships between three distinct dive behaviours observed in lactating northern fur seals to the time of day and measures of walleye pollock abundance around St. Paul Island, Alaska. Our analysis has shown that the interaction of time and prey abundance provides interesting insights into the organization of dive behaviours around a 24 hour, diel pattern.

Northern Fur Seals exhibit a broad foraging range

Our data were collected during the summer pupping seasons when female northern fur seals are nourishing pups and tied to a central place. This limits their effective foraging range in both time spent away, and distance traveled from the natal rookery of their dependent pup. Female northern fur seals in this study embarked on relatively long pelagic journeys averaging 279 km in linear distance from the rookery, and lasting from 5.5 to 11.2 days. This number is consistent but slightly longer than the linear distances reported in Nordstrom et al. 2013 (228 km) and Goebel et al. 1991 (200 km), and the length of time of the foraging trip is consistent with reported lengths from other studies (e.g. Gentry and Holt 1986, Nordstrom et al. 2013; from 1 to 14 days). Predictable food is thought to drive seasonal movements because pup survival depends on the amount of food available to parents for building body reserves prior to provisioning their offspring. Therefore within a broad range of a central place, foraging habitat is selected in which they must get energy to feed themselves and their pup while still returning to the rookery in time to feed a waiting pup.

All of the fur seals showed directional movement, but the foraging tracks of our tagged females were unconstrained and at the broadest scale did not appear collectively to prefer any one region around St. Paul Island. Individuals may, however, have such preferences. The one female for which we had two consecutive tracks went to a similarly located feeding area off-shelf in both foraging trips. Call et al. (2008) similarly showed that 27 of 36 female fur seals with repeated recorded trips tended to travel to the same type of site following the same general direction they used in each of their previous trips. This implies that fur seals know where they are going, and suggests they may adopt a strategy of returning to the same places as used previously to forage for as long as this strategy is successful. However, none

of our tracks, or those recorded by other researchers, showed consensus among the fur seals about where to feed, as none of the tracks showed fur seals traveling to a single location before swimming directly back to nurse their pups. Rather the 3-dimensional dive profiles (e.g. Figure 5.6) and horizontal behaviour maps (Figure 5.9) showed consistent changes in diving behaviours throughout the at-sea foraging trips with active and exploratory dive behaviours distributed over much of the paths they swam. In other words, there was no single grocery store or hot spot where the fur seals spent concentrated time. In other studies, northern fur seals from Reef Rookery have been found to use all hydrographic domains around St. Paul Island (eg. Robson et al. 2004), including both on-shelf and off-shelf habitats (Loughlin et al. 1987, Goebel et al. 1991, Sterling and Ream 2004, Call et al. 2008), and scat samples from the same rookery contained both on-shelf (e.g. walleye pollock, Pacific herring) and off-shelf species (e.g. myctophids; Zeppelin and Orr 2010). The lactating fur seals in this study also appear to be constantly on the move, actively feeding and exploring the ocean as they traveled through broad regions of the Bering Sea.

Persistence

We do not know what the animals we tracked were seeking or what they caught during their foraging trips. However, juvenile walleye pollock of year-classes 0 to 5 are the most common prey in both stomach and scat diet studies (Perez and Bigg 1986, Sinclair et al. 1994, Sinclair et al. 1996, Gudmundson et al. 2006, Zeppelin and Ream 2006), and walleye pollock comprised 89.3% of fur seal diet measured in scats collected within a month of our 2006 tagged females foraging trips from Reef Rookery (Zeppelin and Orr 2010). We found that in at-sea locations with more abundant walleye pollock, northern fur seals were more likely to be actively diving, particularly if those locations overlapped with the nighttime. It is unlikely that our source for the horizontal distribution and abundance of walleye pollock, which was the US Department of Commerce domestic observer data of the Alaska groundfish industry aggregated over both time and space, is directly measuring the real-time abundance of prey that the fur seals actually encountered along their foraging route at any single location. It is more likely that our measure of walleye pollock represents a persistent horizontal feature that is attractive to foraging northern fur seals, but not a very accurate measure of prey availability. Likewise, Nordstrom et al. (2013) linked foraging of northern fur seals to oceanic surface fronts (eddies and filaments) and Kuhn (2011) to thermocline depth; features that may also be proxies for the horizontal and vertical distribution of prey species.

Hence, available evidence indicates that fur seals forage more heavily in areas where either commercial catch data (this study) or oceanographic conditions (Kuhn 2011, Nordstrom et al. 2013) point to increased prey abundance. As well, individuals tend to return to the same general area on successive foraging trips, from where they have no trouble finding the way back to the rookery. Hence, they certainly navigate successfully over hundreds of kilometres in the Bering Sea. Yet little is known to date about the extent to which they can identify potentially persistent foraging "hot-spots" in advance.

Competition with Fisheries

In this study we did not include any direct measure of the bycatch of juvenile pollock in the commercial fishery, nor did we have any small scale evidence that commercial boats were fishing at the same time and location as our tagged females, thus we cannot draw conclusions about the potential for competition occurring between the commercial fishery and northern fur seals. However, we can say that being in the same area where fishing occurred would not necessarily imply competitive interactions unless it was known that the fur seals were feeding on the same fish targeted by the fishery, and whether the abundance of that target prey was limited. Evidence suggests that juvenile pollock of year-class 0 to 5 are preferred by foraging fur seals (Gudmundson et al. 2006, Zeppelin and Ream 2006), which are typically smaller than the > 40 cm adults taken by the commercial fishery (Ianelli et al. 2007). However, inter-year variability of pollock can be as high as two orders of magnitude (1.7 t/km^2 in 2004, 328 t/km^2 in 2007 for the same region; Sigler et al. 2012). In anomalous years with poor juvenile survival combined with poor persistence of the available prey, it could be difficult for foraging females to find prey.

Dive types

Our analysis is built upon the numerical properties of an AR(2) movement model to describe dives in terms of foraging behaviour rather than focusing on the depth at which the dives were, or the tortuosity of that path. The AR(2) process provides a mechanistic way of categorizing dive types, that is based around interpretable movement signatures, and therefore is exempt from arbitrary classification such as those distinguishing dive types based on depth (e.g. Gentry et al. 1986, Goebel 2002). Furthermore, the mechanistic approach protects us from limitations of other studies where deeper dives associated with foraging at or near the

bottom were observed when fur seals were moving in straight lines and not identified as the typical tortuous paths of foraging regions (eg. Nordstrom et al. 2013, Benoit-Bird 2013a, Benoit-Bird 2013b).

Each of the three behavioural modes described by dive types has an intuitive movement interpretation. Shallow, repeated dives with short surface-time intervals, as well as deeper, repeated U-shaped dives with comparatively longer surface recovery times are well described by the solutions of a periodic AR(2) process. U-shaped dives have been attributed to foraging behaviours in an array of top marine predators such as gray seals (Austin et al. 2006), southern elephant seals (Gallon et al. 2013), Australian fur seals (Arnould and Hindell 2001), harbour seals (Baechler et al. 2002), Gentoo Penguins (Wilson et al. 1996), and others. When the surface time between dives is longer but the vertical profile of the dive is more V-shaped, the solutions reflect a correlated random walk without a periodic component. It is less clear what the underlying motivation for intermittent V-shaped dives is, but others have attributed them to foraging on larger pelagic fish or squid in northern gannets (Garthe et al. 2000), or non-foraging activities, including predator avoidance, and explorations (Bengston and Stewart 1992). Non-diving and non-foraging behaviours such as resting, sleeping, grooming and surface transiting are described in the solution space as uncorrelated white noise.

Diel pattern

Modeling results depicted in Figure (5.11) highlight the strength of the diel pattern in foraging northern fur seals, and that the dive pattern the female engages in is that which best fits her prey environment. For example, we found that females showed strong preferences for active diving at night, while preferring non-dive behaviours such as resting or transiting in the mornings. Afternoons were typically associated with exploratory V-shaped dive behaviours. The model suggests that nighttime is characterized by repetitive dives of active feeding (or an AR(2) periodic solution). Kuhn et al. (2010) proposed that shallow foraging bouts at night may be related to juvenile walleye pollock's nighttime migrations to the surface as they follow vertically migrating zooplankton.

It should be energetically advantageous to the northern fur seal to be able to catch prey at shallower depths, resulting in a shorter chase, and possibly also allowing the fur seal to keep physical contact with the school of prey. Northern fur seals may therefore prefer epipelagic (shallow) diving, and transition into deeper U-shaped foraging dives only

when epipelagic resources in an area are scarce. For example, deeper foraging behaviour has been associated with fur seals targeting concentrated groups of juvenile pollock at the thermocline (Kuhn 2011, Nordstrom et al. 2013). If northern fur seals are spatially aware of broad-scale suitable feeding areas based on prior experience or some other knowledge, then it is possible they could reduce search time and energetic costs of foraging by anticipating changes in broad scale prey density and positioning themselves strategically at optimal foraging locations come nightfall.

Data requirements

These eleven fur seal foraging trips generated some useful, basic insight into at-sea behavioural patterns across time and space. The Kullback-Leibler values portrayed in Figure (5.12) suggest that another 10 to 20 tracks would have generated useful further precision. We also anticipate that other oceanographic information on, for example, frontal eddies and filament locations, could provide valuable further insight into the association between fur seals and the Bering Sea environment. Within the hierarchical structure, it would be interesting to link the foraging behaviour characteristics of the maternal fur seal to her pup weight gains, as well as residency times and other life history characteristics that would contribute to our understanding of pup survival.

5.5 Conclusions

Our analysis did not attempt to answer the question of where in the ocean did the fur seals chose to forage relative to not forage. With only 11 tracks over an area with a radius of $> 200\ km$ (or $> 125,000\ km^2$), this would have been too ambitious. Instead we answer the conditional question, "Given that the fur seal swam through this location, can we predict its most likely behaviour based on a set of local environmental variables".

With the hierarchical Bayesian approach used in this study, we were able to link together northern fur seals that went to disparate regions of the eastern Bering Sea, with widely variable information about their underlying environmental fields into a single model that informed us about the expected behaviour of a population of maternal, female northern fur seals at Reef Rookery on St. Paul Island. Our analysis has focussed on three typical fur seal behaviours, and how these behaviours are associated with time of day, and a set of environmental data. The approach and analysis presented here is to use hierarchical

Bayesian modelling to bring together coarse estimates of the at-sea environment, and the dynamics of fur seal behaviour through identification of varaible-length segments of coherent behaviours. In this, we have successfully shown that northern fur seal behaviour actively forage more at night, and in areas where their preferred prey species, juvenile walleye pollock, appears to be more abundant.

5.6 Literature Cited

Akaike, H. 1973. Information theory and an extension of the maximum likelihood principle. Second International Symposium on Information Theory, Academiai Kiado, Budapest, pp. 267-281.

Amante, C., and B.W. Eakins. 2009. ETOPO1 1 Arc-Minute Global Relief Model: procedures, data sources and analysis. NOAA Technical Memorandum NESDIS NGDC-24, 19pp.

Antonelis, G.A., E.H. Sinclair, R.R. Ream, and B.W. Robson. 1997. Inter-island variability in the diet of female northern fur seals (*Callorhinus ursinus*) in the Bering Sea. Journal of Zoology, London, 242: 435-451.

Arnould, J.P.Y. and M.A. Hindell. 2001. Dive behaviour, foraging locations, and maternal attendance patterns of Australian fur seals (*Arctocephalus pusillus doriferus*). Canadian Journal of Zoology, 79: 35-48.

Austin, D.A., W.D. Bowen, J.I. McMillan, and S.J. Iverson. 2006. Linking movement, diving, and habitat to foraging success in a large marine predator. Ecology, 87(12): 3095-3108.

Baechler, J., C.A. Beck, and W.D. Bowen. 2002. Dive shapes reveal temporal changes in the foraging behaviour of different age and sex classes of harbour seals (*Phoca vitulina*). Canadian Journal of Zoology, 80: 1569-1577.

Bayarri, M.J. and J.O. Berger. 2000. *P*-values for composite null models. Journal of the American Statistical Association, 95(452): 1127-1142.

Behrenfeld, M.J. and P.G. Galkowski. 1997. Photosynthetic rates derived from satellite-based chlorophyll concentration. Limnology and Oceanography, 42: 1-20, doi:10.4319/lo.1997.42.1.0001

Bengston, J.L., R.L. Merrick, and T.R. Loughlin. 1985. Radio-tracking studies, St. Paul Island, Alaska. *In* Fur seal investigations, 1985. Edited by P. Kozloff and H. Kajimura. NOAA Tech. Memo., NMFS F/NWC-146. Pp. 104-106.

Benoit-Bird, K. J., Battaile, B. C., Nordstrom, C. A., and Trites, A. W. 2013. Foraging behavior of northern fur seals closely matches the hierarchical patch scales of prey. Marine Ecology Progress Series, 479, 283-302.

Benoit-Bird, K.J., B.C. Battaile, S.A. Heppell, B. Hoover, D. Irons, N. Jones, K.J. Kuletz, C.A. Nordstrom, R.Paredes, R.M. Suryan, C.M. Waluk, and A.W. Trites, 2013. Prey Patch Patterns Predict Habitat Use by Top Marine Predators with Diverse Foraging Strategies. PloS ONE, 8(1), e53348.

Berkson, J. 1950. Are there two regressions? Journal of the American Statistical Association, 45(250): 164-180.

Bowler, D.E. and T.G. Benton. 2005. Causes and consequences of animal dispersal strategies: relating individual behaviour to spatial dynamics Biological Reviews, 80: 205-225.

Boyd, I.L. and Croxall, J.P. 1992: Diving behaviour of lactating Antarctic fur seals. Canadian Journal of Zoology, 70: 919-928.

Brooks, S.P., and A. Gelman. 1998. General methods for monitoring convergence of iterative simulations. Journal of Computational and Graphical Statistics, 7(4): 434-455.

Burnham, K.P., and D.R. Anderson. 2001. Kullback-Leibler information as a basis for strong inference in ecological studies. Wildlife Research. 28: 111-119.

Call, K.A., R.R. Ream, D. Johnson, J.T. Sterling, and R.G. Towell. 2008. Foraging route tactics and site fidelity of adult female northern fur seal (*Callorhinus ursinus*) around the Pribilof Islands. Deep-Sea Research Part II. 55(16): 1883-1896.

Carlin, B., and T.A. Louis. 2000. Bayes and empirical Bayes methods for data analysis (2^{nd} ed.). 418 pp.

Clyde, M.A. 2000. Model uncertainty and health effect studies for particulate matter. Environmetrics, 11:745:763.

Cressie, N., C.A. Calder, J.S. Clark, J.M. Ver Hoef, and C.K. WIkle. 2009. Accounting for uncertainty in ecological analysis: the strengths and limitations of hierarchical statistical modelling. Ecological Applications, 19(3): 553-570.

Dempster, A.P. 1974. The direct use of likelihood for significance testing. *In* Proceedings of Conference on Foundational questions in statistical inferences, (*ed.* O. Barndorff-Nielsen, P. Blaesild, and G. Schou), pp.335-352. Department of Theoretical Statistics: University of Aarhus.

Gallon, S., F. Bailleul, J.-B. Charrassin, C. Guinet, C.-A. Bost, Y. Handrich and M. Hindell. 2013. Identifying foraging events in deep diving southern elephant seals, *Mirounga leonina*, using acceleration data loggers. Deep-Sea Research II, 88-89: 14-22.

Garthe, S., S. Benvenuti and W.A. Montevecchi. 2000. Pursuit plunging by northern gannets (*Sula bassana*) "feeding on capelin (*Mallotus villosus*)". Proceedings of the Royal Society B, 267(1454): 1717-1722.

Gelman, A., J.B. Carlin, H.S. Stern, and D.B. Rubin. 2004. Bayesian data analysis (2^{nd} ed.). Boca Raton, FL: CRC Press. 668 pp.

Gentry, R.L., and J.R. Holt. 1986. Attendance behavior of northern fur seals. *In* Fur seals, maternal strategies on land and at sea. Edited by R. L. Gentry and G. L. Kooyman. Princeton University Press, Princeton, NJ, pp. 41-60.

Gentry, R.L., G.L. Kooyman, and M.E. Goebel. 1986. Feeding and diving behavior of northern fur seals. Pp. 61-78. *In* R. L. Gentry and G. L. Kooyman (*eds.*). Fur seals: maternal strategies on land and at sea, Princeton University Press, Princeton.

Goebel, M., J. Bengtson, R. DeLong, R. Gentry, and T.R. Loughlin. 1991. Diving patterns and foraging locations of female northern fur seals. Fishery Bulletin, 89: 171-179.

Goebel, M.E. 2002. Northern fur seal lactation, attendance and reproductive success in two years of contrasting oceanography. PhD thesis, University of California, Santa Cruz. 227 pp.

Gudmundson, C.J., T.K. Zeppelin, and R.R. Ream. 2006. Application of two methods for determining diet of northern fur seals (Callorhinus ursinus). Fishery Bulletin, 104: 445-455.

Harcourt, R., and L. Davis. 1997. The use of satellite telemetry to determine fur seal foraging areas. Pp. 137-142. *In* M. Hindell and C. Kemper (eds.). Marine Mammal Research in the Southern Hemisphere Volume 1: Status, Ecology, and Medicine, Surrey Beatty & Sons.

Herrmann, E. 2003. Lokern: An R package for kernel smoothing.

Johnson, V. 2004. A Bayesian χ^2-test for goodness of t. Annals of Statistics, 32(6): 2361-2384.

Johnson, J.B. and K.S. Omland. 2004. Model selection in ecology and evolution. Trends in ecology & evolution, 19(2): 101-108.

Kass, R.E., B.P. Carlin, A. Gelman, and R.M. Neal. 1998. Markov chain monte carlo in practice: A roundtable discussion. The American Statistician, 52(2): 93-100.

Kuhn, C.E., Y. Tremblay, R.R. Ream, and T.S. Gelatt. 2010. Coupling GPS tracking with dive behavior to examine the relationship between foraging strategy and ne-scale movements in northern fur seals (*Callorhinus ursinus*). Endangered Species Research, 12:125-139.

Kuhn, C.E. 2011. The inuence of subsurface thermal structure on the diving behavior of northern fur seals (*Callorhinus ursinus*) during the breeding season. Marine Biology, 158: 649-663.

Kullback, S., and R.A. Leibler. 1951. On information and sufficiency. The Annals of the Institute of Statistical Mathematics, 22: 79-86.

Lee., Y.K., and B.U. Park. 2006. Estimation of Kullback-Leibler divergence by local likelihood. The Annals of the Institute of Statistical Mathematics, 58(2): 326-340.

Leisch, F. 2004. FlexMix: A General Framework for Finite Mixture Models and Latent Class Regression in R. Journal of Statistical Software, 11(8): 1-18.

Loughlin, T.R., J.L. Bengtson, and R.L. Merrick. 1987. Characteristics of feeding trips of female northern fur seals. Canadian Journal of Zoology, 65: 2079-2084.

Loughlin, T.R., I.N. Sukhanova, E.H. Sinclair, and R.C. Ferrero. 1999. Summary of biology and ecosystem dynamics in the Bering Sea. Dynamics of the Bering Sea, 99(1999): 387407.

Meng, X.L. 1994. Posterior predictive p-values. The Annals of Statistics, 22(3): 1142-1160.

Mitani, Y., K. Sato, I. Shinichiro, M.F. Cameron, D.B. Siniff, and Y. Naito. 2003. A method for reconstructing three-dimensional dive profiles of marine animals using geomagnetic intensity data: results from two lactating Weddell seals. Polar Biology, 26: 311-317.

NMFS. 2007. Conservation plan for the Eastern Pacific stock of northern fur seal (*Callorhinus ursinus*). National Marine Fisheries Service, Juneau, Alaska.

NMFS. 2012. National Observer Program Annual Report FY 2011, NOAA Tech. Memo. NMFS F/SPO-123, 36 pp.

Nordstrom, C.A., B.C. Battaile, C. Cotté, and A.W. Trites. 2013. Foraging habitats of lactating northern fur seals are structured by thermocline depths and submesoscale fronts in the eastern Bering Sea. Deep-Sea Research II, http://dx.doi.org/10.1016/j.dsr2.2012.07.010

Perez, M.A., and M.A. Bigg. 1986. Diet of northern fur seals, *Callorhinus ursinus*, off western North America. Fishery Bulletin, 84: 957-971.

Pitt, M.A., and I.J. Myung. 2002. When a good fit can be bad. Trends in Cognitive Science, 6: 421-425.

Ream, R.R., J.T. Sterling, and T.R. Loughlin. 2005. Oceanographic features related to northern fur seal migratory movements. Deep Sea Research Part II: Topical Studies in Oceanography, 52(5), 823-843.

Robson, B. W., M. E. Goebel, J. D. Baker, R.R. Ream, T.R. Loughlin, R.C. Francis, G.A. Antonelis, and D.P.Costa. 2004. Separation of foraging habitat among breeding sites of a colonial marine predation, the northern fur seal (*Callorhinus ursinus*). Canadian Journal of Zoology, 82: 2029.

Rubin, D.B. 1984. Bayesian justifiable and relevant frequency calculations for the applied statistician. The Annals of Statistics, 12: 1152-1172.

Sinclair, E.H., T.R. Loughlin, and W. Pearcy. 1994. Prey selection by northern fur seals (*Callorhinus ursinus*) in the eastern Bering Sea. Fishery Bulletin, 92: 132-156.

Sinclair, E.H., G.A. Antonelis, B.W. Robson, R. Ream, and T.R. Loughlin. 1996. Northern fur seal, *Callorhinus ursinus*, predation on juvenile pollock, *Theragra chalcogramma*. Pp. 167-178. *In* R. D. Brodeur, P. A. Livingston, T. R. Loughlin, and A. B. Hollowed, (*eds.*). Ecology of juvenile walleye pollock, *Theragra chalcogramma*: papers from the workshop "The importance of pre-recruit walleye pollock to the Bering Sea and North Pacific ecosystems" Seattle, Washington, 28-30 October 1993. U.S. Department of Commerce, NOAA Technical. Report. NMFS-126. 227 pp.

Sigler, M.F., K.J. Kuletz, P.H. Ressler, N.A. Friday, C.D. Wilson, and A.N. Zerbini. 2012. Marine predators and persistent prey in the southeast Bering Sea. Deep-Sea Research II, 65-70: 292-303.

Spiegelhalter, D., A. Thomas, N. Best, and D. Lunn. 2002. WinBUGS User Manual Version 1.4. Medical Research Council Biostatistics Unit: Cambridge, UK.

Sterling, J.T., and R.R. Ream. 2004. At-sea behavior of juvenile male northern fur seals (*Callorhinus ursinus*). Canadian Journal of Zoology, 82: 1621-1637.

Towell, R.G., R.R. Ream, and A.E. York. 2006. Decline in northern fur seal (*Callorhinus ursinus*) pup production on the Pribilof Islands. Marine Mammal Science, 22(2): 486-491.

Wilson, R.P., B.M. Culik, G. Peters, R. Bannasch. 1996. Diving behaviour of Gentoo penguins, *Pygoscelis papua*; factors keeping dive profiles in shape. Marine Biology, 126: 153-162.

Zeppelin, T.K., and R.R. Ream. 2006. Foraging habitats based on the diet of female northern fur seals (*Callorhinus ursinus*) on the Pribilof Islands, Alaska. Journal of Zoology, 270: 565-576.

Zeppelin, T.K., and A.J. Orr. 2010. Stable isotope and scat analyses indicate diet and habitat partitioning in northern fur seals *Callorhinus ursinus* across the eastern Pacific. Marine Ecology Progress Series, 409: 241-253.

Zhang, H.M., J.J. Bates, and R.W. Reynolds. 2006. Assessment of composite global sampling: Sea surface wind speed. Geophysical Research Letters, 33(17).

Chapter 6

Conclusions

This thesis considers state space estimation techniques, and hierarchical Bayesian modelling, and develops behavioural insights about northern fur seals foraging in the Bering Sea using these models. The thesis embodies three research contributions as well as a discussion of extensions of these. The first contribution is to provide a statistical framework from which continuous parameter estimates that are interpretable in terms of animal movement can be extracted. The second contribution develops a multinomial hierarchical model framework in which coherent behaviours can be identified and extracted from the continuous movement state space solutions and linked to course measures of environmental covariates. The third is an ecological application that relates changes in three identifiable behaviours to changes in catch in the Alaska commercial groundfish industry over a circadian foraging cycle.

In this thesis, we explore the idea of estimating behaviour parameters from movement models of marine animal archival tag data within a state space framework. The central idea is that by estimating the time variation of parameters for a suitable movement model, we can then objectively and quantitatively infer animal activity, and its behavioural state. Here, we offer a statistical-dynamical approach suitable for extracting behavioural information from high resolution data, and one that is widely applicable to a variety of movement models and observation types. It offers an alternative to behavioural switching state space models (Jonsen et al. 2007, Patterson et al. 2009), and is flexible enough to obtain solutions even for nonlinear and non-Gaussian cases. Our approach draws from the state augmentation procedures of Kitagawa (1998) and Ionides et al. (2006).

Our application focusses on analysis and interpretation of the vertical movement data from a tagged northern fur seal. The analysis provides slowly time-varying estimates for

the movement parameters.

The second objective is centered on using the fine scale behavioral information from the state space model to construct a model linked to the fur seal's physical and biological environment. We explore building a two-level hierarchical Bayesian model to draw inference about animal behaviour. One important aspect of the process error in hierarchical models is the difficulty in accurately estimating parameters of interest when there is error contained within the covariate data. Many of the spatial covariates available for analysis in marine mammal tracking problems remain constant over loose grids, or large sections of the ocean. The hierarchical modelling framework is an ideal means to account for sources of uncertainty in the measurement and process models that are inherent in these kinds of ecological settings. These methods that we propose are one of a set of emerging flexible, analytical techniques that consider animal movement to be three-dimensional, and can, with enough data, take any number of levels in the hierarchy, while accounting for multiple sources of uncertainty.

The final objective is to reconstruct a set of high-resolution three-dimensional northern fur seal foraging tracks using positional data, depth data and velocity data, and link these to the physical and biological environment of the eastern Bering Sea. Using these at-sea tracks, we link together northern fur seals that went to disparate regions of the eastern Bering Sea and create a single model that informs us about the expected behaviour of the population of maternal, female northern fur seals at Reef Rookery, on St. Paul Island. We show that northern fur seals actively forage more at night, and in areas where their preferred prey species, juvenile walleye pollock, appears to be more abundant. This understanding comes at a key stage in the life of northern fur seals when foraging success is extremely important for pup survival.

These results describe nursing female northern fur seals at Reef Rookery. However Call and Ream (2012) found high levels of dietary overlap among subadult male and adult female fur seals, thus it would be interesting to explore if the diel relationship between behaviour and juvenile walleye pollock found in this study, extends to subadult males.

While the primary focus of this thesis has been on northern fur seals, it is important to emphasize that the methodology is most definitely applicable to analyzing any type of high-resolution movement data on not just marine mammals, but for example, seabirds, or other marine predators. The hierarchical Bayesian model framework will allow for the characterization of the relationship of the behaviours we classified, but also of other behaviours as well as with other descriptors of pelagic habitat and foraging success.

Recent review articles have identified state space models (Patterson et al. 2008) and hierarchical Bayesian approaches (Schick et al. 2008) as two important directions for extracting ecologically meaningful information from animal tag data. Here we have used a particle filter approach for an augmented state space model tailored to the type of large volume, high resolution time series recorded by archival animal tags. As well, we have used a hierarchical Bayesian approach to link the output of the state space models to an animal's at-sea environment. The approach of this study is a promising direction for fully exploiting the behavioural information in these rich data sets.

6.1 Literature Cited

Call, K.A., and R.R. Ream. 2012. Prey selection of subadult male northern fur seals (*Callorhinus ursinus*) and evidence of dietary niche overlap with adult females during the breeding season. Marine Mammal Science, 28(1): 1-15.

Ionides, E.L., C. Breto, and A.A. King. 2006. Inference for nonlinear dynamical systems. Proceedings of the National Academy of Sciences, 103: 18438-18443.

Jonsen, I.D., R.A. Myers, and M.C. James. 2007. Identifying leatherback turtle foraging behaviour from satellite-telemetry using a switching state space model. Marine Ecology Progress Series, 337: 255-264.

Kitagawa, G. 1998. A self-organising state space model. Journal of the American Statistical Association, 93: 1203-1215.

Patterson, T.A., L. Thomas, C. Wilcox, O. Ovaskainen, and J. Matthiopoulos. 2008. State-space models of individual animal movement. Trends in Ecology and Evolution, 23: 87-94.

Patterson, T.A., M. Basson, M.V. Bravington, and J.S. Gunn. 2009. Classifying movement behaviour in relation to environmental conditions using hidden Markov models. Journal of Animal Ecology, 78: 1113-1123.

Schick, R.S., S.R. Loarie, F. Colchero, B.D. Best, A. Boustany, D.A. Conde, P.N. Halpin, L.N. Joppa, C.M. McClellan, and J.S. Clark. 2008. Understanding movement data and movement processes: current and emerging directions. Ecology Letters, 11: 1338-1350.

Appendix A

Specifying Errors in Chapter 3

Determining the observation error variance σ_o^2 and the system noise variance σ_ϵ^2 is an important part of the state space model. Note that in principle σ_ϵ^2 and σ_o^2 could be estimated as part of the parameter estimation procedure using the state augmentation approach. However, it is often difficult to separately estimate these quantities using a state space model, as they may be confounded with each other, or worse with the movement parameters. It is therefore advisable to estimate any parameters off-line (where possible) to ensure identifiability of the movement parameters. Such a procedure is outlined below for our application, and is used to specify the observation error variance in (Equation 3.1), as well as the magnitude of the system noise variance, which scales the normal mixture model (Equation 3.12).

The calculation of the sample auto-covariance (ACVF) function provides the basis for our estimation procedure. We assume that the observation error is uncorrelated through time, and hence that the observation error variance contributes to the ACVF only at zero lag, and has zero auto-covariance for all other lags. In contrast, the state vector for vertical velocity, x_t, is correlated through time as dictated by the state evolution equation (Equation 3.9), and so has contributions to the ACVF at both zero and non-zero lags.

Based on the above assumptions, the theoretical ACVF, designated as $\gamma(k)$, for observed vertical velocity, y_t, is given by

$$\gamma(k) = \sigma_o^2(0) + \sigma_x^2(k) \tag{A.1}$$

where k denotes the lag. At zero lag the total variance, $\gamma(0)$, has contributions from observation error variance, $\sigma_o^2(0)$, and the process variance, $\sigma_x^2(0)$; at non-zero lags it has

contributions only from $\sigma_x^2(k)$.

The sample ACVF, $\hat{\gamma}(k)$ of Figure (3.1) can therefore be used to separate σ_o^2 from σ_x^2 using (A.1). To do this, a quadratic was fit to $\hat{\gamma}(k)$ associated with small non-zero lags $k = 1, \ldots n$. We chose $n = 8$ to capture the quadratic trend near the origin. After fitting we then extrapolate to the origin, $k = 0$, to provide an estimate of the process variance at lag 0, or $\hat{\sigma}_x^2(0)$. Based on this, the estimated observation error variance is taken as $\hat{\sigma}_o^2 = \hat{\gamma}(0) - \hat{\sigma}_x^2(0)$. Note that this procedure is directly analogous to determining the nugget in kriging.

We next determine the system noise variance, σ_ϵ^2, using our estimate of the process noise variance, σ_x^2. If we assume that x_t follows a Markov process, or has an exponential correlation function, then the system noise variance can be estimated as (Priestley 2004, Section 3.5.3),

$$\hat{\sigma}_\epsilon^2 = \left(1 - \left(\frac{\hat{\gamma}(1)}{\hat{\gamma}(0)}\right)^2\right) \hat{\sigma}_x^2(0).$$

This is an approximation and relies on the ratio of sample ACVF at lag 1 to lag 0 to describe the decay rate of the underlying autocovariance function, but acts as a variance adjustment due to the autocorrelation in our data, while making no assumptions about the movement process itself.

Both σ_o^2 and σ_ϵ^2 change over the dive record. The evolutionary sample ACVF [Figure 3.3 (i)] was used with the above approach to compute the time evolution of the observation error variance and the system noise variance for each of our 110 time windows. These estimates are shown in Figure [3.4 (i)]. The resultant values obtained for $\sigma_{\epsilon,t}^2$ and $\sigma_{o,t}^2$ can be used directly as input to the state space model.

A.1 Literature Cited

Priestley, M.B. 2004. Spectral Analysis and Time Series. Academic Press. London. 890pp.

Appendix B

Target Distributions of the Multinomial Model in Chapter 4

B.1 Likelihood

$$P(\boldsymbol{y}_i|\boldsymbol{\beta}_i, \boldsymbol{X}_i, \boldsymbol{Z}_i) \propto exp\left\{ \sum_j \boldsymbol{y}_{ij} \left(\boldsymbol{\beta}_{iZ}\boldsymbol{Z}_{ij} + \boldsymbol{\beta}_{iX}\boldsymbol{X}_{ij}\right) - \sum_j \log\left(1 + e^{\boldsymbol{\beta}_{iZ}\boldsymbol{Z}_{ij} + \boldsymbol{\beta}_{iX}\boldsymbol{X}_{ij}}\right) \right\}.$$

B.2 Prior Distributions

$$\begin{aligned}
&\boldsymbol{\beta}_i \sim \mathcal{MVN}(\mathbf{1}\boldsymbol{B}, \boldsymbol{\Sigma}_\beta) \\
&\boldsymbol{X}_i \sim \mathcal{MVN}(\boldsymbol{W}_i, \sigma^2_{iU}) \\
&\sigma^2_{iU} \sim \mathcal{IG}(a_U, b_U) \\
&\boldsymbol{\Sigma}_\beta \sim \mathcal{IW}(\nu_0, \boldsymbol{V}_0) \\
&\boldsymbol{B} \sim \mathcal{MVN}(\boldsymbol{B}_0, \frac{1}{a_0}\boldsymbol{\Sigma}_\beta)
\end{aligned}$$

B.3 Posterior Distributions

B.3.1 Joint Posterior Distribution

$$
\begin{aligned}
P(\boldsymbol{\beta}, \boldsymbol{X}, \sigma_U^2, \boldsymbol{B}, \boldsymbol{\Sigma}_\beta | \boldsymbol{y}_i, \boldsymbol{Z}_i, \cdot) \propto \prod_{i=1}^{n} & \Bigg(P(\boldsymbol{y}_i | \boldsymbol{X}_i, \boldsymbol{\beta}_i, \boldsymbol{Z}_i) P(\boldsymbol{X}_i | \boldsymbol{W}_i, \sigma_{iU}^2) \times \\
& P(\boldsymbol{\beta}_i | \boldsymbol{B}, \boldsymbol{\Sigma}_\beta) P(\sigma_{iU}^2 | a_U, b_U) \Bigg) \times \\
& P(\boldsymbol{B}, \boldsymbol{\Sigma}_\beta | \nu_0, \boldsymbol{V}_0, \boldsymbol{B}_0, \frac{1}{a_0}) \\
\propto \prod_{i=1}^{n} & \Bigg(exp \left\{ \sum_{j=1}^{m_i} \boldsymbol{y}_{ij} \left(\boldsymbol{\beta}_{iZ} \boldsymbol{Z}_{ij} + \boldsymbol{\beta}_{iX} \boldsymbol{X}_{ij} \right) - \right. \\
& \left. \sum_{j=1}^{m_i} \log \left(1 + e^{\boldsymbol{\beta}_{iZ} \boldsymbol{Z}_{ij} + \boldsymbol{\beta}_{iX} \boldsymbol{X}_{ij}} \right) \right\} \times \\
& \frac{1}{\sqrt{\left(2\pi\sigma_{iU}^2 \right)^{m_i}}} exp \left\{ (2\sigma_{iU}^2)^{-1} \sum_{j=1}^{m_i} (\boldsymbol{X}_{ij} - \boldsymbol{W}_{ij})^2 \right\} \times \\
& |\boldsymbol{\Sigma}_\beta|^{-n/2} exp \left\{ -\frac{1}{2} (\boldsymbol{\beta}_i - \mathbf{1}\boldsymbol{B}) \, \boldsymbol{\Sigma}_\beta^{-1} \, (\boldsymbol{\beta}_i - \mathbf{1}\boldsymbol{B}) \right\} \times \\
& (\sigma_{iU}^2)^{a_U + 1} exp \left\{ -b_U / \sigma_{iU}^2 \right\} \Bigg) \times \\
& |\boldsymbol{\Sigma}_\beta|^{-(\nu_0 + p + 1)/2} exp \left\{ -\frac{1}{2} (\boldsymbol{V}_0 \boldsymbol{\Sigma}_\beta^{-1}) \right\} \times \\
& |\boldsymbol{\Sigma}_\beta|^{-1/2} exp \left\{ -\frac{1}{2} \Big((\boldsymbol{B} - \boldsymbol{B}_0) \boldsymbol{A} \boldsymbol{\Sigma}_\beta^{-1} (\boldsymbol{B} - \boldsymbol{B}_0) \Big) \right\}.
\end{aligned}
$$

B.3.2 Lower-level posterior distributions

Marginal Posterior Distribution of $\boldsymbol{\beta}$

$$\begin{aligned} P(\boldsymbol{\beta}_i|\, \boldsymbol{y}_i, \boldsymbol{X}_i, \boldsymbol{Z}_i, \boldsymbol{B}, \boldsymbol{\Sigma}_\beta) \quad \propto \quad & exp\left\{ \sum_{j=1}^{m_i} \boldsymbol{y}_{ij} \left(\boldsymbol{\beta}_{iZ}\boldsymbol{Z}_{ij} + \boldsymbol{\beta}_{iX}\boldsymbol{X}_{ij}\right) - \right. \\ & \left. \sum_{j=1}^{m_i} \log\left(1 + e^{\boldsymbol{\beta}_{iZ}\boldsymbol{Z}_{ij} + \boldsymbol{\beta}_{iX}\boldsymbol{X}_{ij}}\right) \right\} \times \\ & |\boldsymbol{\Sigma}_\beta|^{-n/2} exp\left\{ -\frac{1}{2}(\boldsymbol{\beta}_i - \mathbf{1}\boldsymbol{B})\, \boldsymbol{\Sigma}_\beta^{-1}\, (\boldsymbol{\beta}_i - \mathbf{1}\boldsymbol{B}) \right\} \end{aligned}$$

Marginal Posterior Distribution of covariate $\boldsymbol{X}_i$

$$\begin{aligned} P(\boldsymbol{X}_i|\, \boldsymbol{y}_i, \boldsymbol{W}_i, \boldsymbol{Z}_i, \boldsymbol{\beta}_i, \sigma^2_{iU}) \propto \qquad & exp\left\{ \sum_{j=1}^{m_i} \boldsymbol{y}_{ij} \left(\boldsymbol{\beta}_{iZ}\boldsymbol{Z}_{ij} + \boldsymbol{\beta}_{iX}\boldsymbol{X}_{ij}\right) - \right. \\ & \left. \sum_{j=1}^{m_i} \log\left(1 + e^{\boldsymbol{\beta}_{iZ}\boldsymbol{Z}_{ij} + \boldsymbol{\beta}_{iX}\boldsymbol{X}_{ij}}\right) \right\} \times \\ & \frac{1}{\sqrt{\left(2\pi\sigma^2_{iU}\right)^{m_i}}}\; exp\left\{ (2\sigma^2_{iU})^{-1} \sum_{j=1}^{m_i} (\boldsymbol{X}_{ij} - \boldsymbol{W}_{ij})^2 \right\} \end{aligned} \tag{B.1}$$

Marginal Posterior Distribution of $\boldsymbol{\sigma}^2_{iU}$, applying $\mathcal{IG}(a_U, b_U)$ prior

$$\begin{aligned} P(\sigma^2_{iU}|\boldsymbol{X}_i, \boldsymbol{W}_i, a_U, b_U) \qquad & \propto \frac{1}{\sqrt{\left(2\pi\sigma^2_{iU}\right)^{m_i}}} exp\{(2\sigma^2_{iU})^{-1} \sum_j (\boldsymbol{X}_{ij} - \boldsymbol{W}_{ij})^2\} \times \\ & (\sigma^2_{iU})^{a_U+1} exp\left\{-b_U/\sigma^2_{iU}\right\} \\ & \propto exp\left\{ \frac{1}{\sigma^2_{iU}} \left[b_U + 1/2 \sum_j (\boldsymbol{X}_{ij} - \boldsymbol{W}_{ij})^2 \right] \right\} (\sigma^2_{iU})^{a_U + \frac{m_i}{2} + 1} \end{aligned} \tag{B.2}$$

$$P(\sigma^2_{iU}|\boldsymbol{X}_i, \boldsymbol{W}_i, a_U, b_U) \sim \mathcal{IG}\left(a_U + \frac{m_i}{2}, \;\; b_U + \frac{1}{2}\sum_{j=1}^{m_i} (\boldsymbol{X}_{ij} - \boldsymbol{W}_{ij})^2 \right).$$

B.3.3 Higher-level conditional posterior distributions

The higher-level priors $\boldsymbol{B}, \boldsymbol{\Sigma}_\beta$ are no longer a function of the data $\boldsymbol{y}, \boldsymbol{X}, \boldsymbol{Z}$, but instead a function of the lower-level parameters $\boldsymbol{\beta}_i$, which are sufficient for the original data. Thus the prior for $\boldsymbol{\beta}_i$ becomes the likelihood of the higher level parameters, and since the dimension of $\boldsymbol{B}$ is $q \times p$, and in our application $q = 1$ and $\boldsymbol{\mathcal{X}} = \mathbf{1}$ (the $n \times 1$ vector of all ones), we can write this likelihood as follows

$$P(\boldsymbol{\beta}_1, \ldots, \boldsymbol{\beta}_{11} | \boldsymbol{B}, \boldsymbol{\Sigma}_\beta, \mathbf{1}) \propto |\boldsymbol{\Sigma}_\beta|^{-n/2} exp\{-\frac{1}{2}(\boldsymbol{\beta}_i - \mathbf{1}\boldsymbol{B})' \, (\boldsymbol{\beta}_i - \mathbf{1}\boldsymbol{B})\boldsymbol{\Sigma}_\beta^{-1}\}. \tag{B.3}$$

To suggest the form of the natural conjugate priors for $\boldsymbol{\Sigma}_\beta$ and $\boldsymbol{B}$, we factor the likelihood (B.3) into two parts

$$\begin{aligned} P(\boldsymbol{\beta}|\boldsymbol{B}, \boldsymbol{\Sigma}_\beta, \mathbf{1}) &\propto |\Sigma_\beta|^{-n/2} exp\left\{-\frac{1}{2}(\boldsymbol{\beta}_i - \mathbf{1}\bar{\boldsymbol{\beta}})'(\boldsymbol{\beta}_i - \mathbf{1}\bar{\boldsymbol{\beta}}) + (\boldsymbol{B} - \bar{\boldsymbol{\beta}})'\mathbf{1}'\mathbf{1}(\boldsymbol{B} - \bar{\boldsymbol{\beta}})\boldsymbol{\Sigma}_\beta^{-1}\right\} \\ &\propto |\Sigma_\beta|^{-(n-1)/2} exp\left\{-\frac{1}{2}(\boldsymbol{\beta}_i - \mathbf{1}\bar{\boldsymbol{\beta}})'(\boldsymbol{\beta}_i - \mathbf{1}\bar{\boldsymbol{\beta}})\Sigma_\beta^{-1}\right\} \times \\ &\quad |\Sigma_\beta|^{-1/2} exp\left\{-\frac{1}{2}(\boldsymbol{B} - \bar{\boldsymbol{\beta}})'\mathbf{1}'\mathbf{1}(\boldsymbol{B} - \bar{\boldsymbol{\beta}})\Sigma_\beta^{-1}\right\} \end{aligned}$$

where $\bar{\boldsymbol{\beta}} = (\mathbf{1}'\mathbf{1})^{-1}\mathbf{1}'\boldsymbol{\beta}$, or the p-column averages of the $\boldsymbol{\beta}$ matrix, and substituting $\boldsymbol{S} = (\boldsymbol{\beta}_i - \mathbf{1}\bar{\boldsymbol{\beta}})'(\boldsymbol{\beta}_i - \mathbf{1}\bar{\boldsymbol{\beta}})$ gives the factored likelihood :

$$\begin{aligned} P(\boldsymbol{\beta}|\boldsymbol{B}, \boldsymbol{\Sigma}_\beta, \mathbf{1}) &\propto |\Sigma_\beta|^{-(n-1)/2} exp\left\{-\frac{1}{2}\boldsymbol{S}\Sigma_B^{-1}\right\} \times \\ &\quad |\Sigma_\beta|^{-1/2} exp\left\{-\frac{1}{2}(\boldsymbol{B} - \bar{\boldsymbol{\beta}})'\mathbf{1}'\mathbf{1}(\boldsymbol{B} - \bar{\boldsymbol{\beta}})\Sigma_\beta^{-1}\right\}. \end{aligned} \tag{B.4}$$

Marginal Posterior Distribution of $\boldsymbol{\Sigma}_\beta$ and $\boldsymbol{B}$

The first line of the factored likelihood in Equation (B.4), suggests an Inverse Wishart ($\mathcal{IW}$) kernel for the covariance matrix $\boldsymbol{\Sigma}_\beta$, and the second line suggests a multivariate normal kernel ($\mathcal{MVN}$) for $\boldsymbol{B}|\boldsymbol{\Sigma}_\beta$. The prior on $\boldsymbol{B}$ is dependent on the scale parameter as $\boldsymbol{\Sigma}_\beta$ cannot be factored out of (B.4), thus the prior for $\boldsymbol{B}$ is a conditionally conjugate prior $P(\boldsymbol{B}|\boldsymbol{\Sigma}_\beta)$. The conjugate $\mathcal{IW}$ prior with hyperparameters $\nu_0, \boldsymbol{V}_0$ was selected for $\boldsymbol{\Sigma}_\beta$, and the conditional conjugate $\mathcal{MVN}$ with hyperparameters $\boldsymbol{B}_0, \frac{1}{a}\boldsymbol{\Sigma}_\beta$ were selected for $\boldsymbol{B}|\boldsymbol{\Sigma}_\beta$ as follows

$$\begin{aligned} &\boldsymbol{\Sigma}_\beta \sim \mathcal{IW}(\nu_0, \boldsymbol{V}_0) \\ &\boldsymbol{B}|\boldsymbol{\Sigma}_\beta \sim \mathcal{MVN}(\boldsymbol{B}_0, \frac{1}{a_0}\boldsymbol{\Sigma}_\beta). \end{aligned}$$

The posterior distributions of $\boldsymbol{B}, \boldsymbol{\Sigma}_\beta$ are then derived

$$\begin{aligned} P(\boldsymbol{B}, \boldsymbol{\Sigma}_\beta|\cdot) &\propto |\boldsymbol{\Sigma}_\beta|^{-n/2} exp\left\{-\frac{1}{2}(\boldsymbol{\beta}_i - \mathbf{1}\boldsymbol{B})'\ (\boldsymbol{\beta}_i - \mathbf{1}\boldsymbol{B})\boldsymbol{\Sigma}_\beta{}^{-1}\right\} \times \\ &\quad |\boldsymbol{\Sigma}_\beta|^{-(\nu_0+p+1)/2} exp\left\{-\frac{1}{2}\boldsymbol{V}_0\boldsymbol{\Sigma}_\beta^{-1}\right\} \times \\ &\quad |\boldsymbol{\Sigma}_\beta|^{-p/2} exp\left\{-\frac{1}{2}(\boldsymbol{B} - \boldsymbol{B}_0)'(\boldsymbol{B} - \boldsymbol{B}_0)\left(\frac{1}{a_0}\boldsymbol{\Sigma}_\beta\right)^{-1}\right\} \end{aligned} \tag{B.5}$$

We combine the terms involving $\boldsymbol{B}$, letting $\boldsymbol{\mathcal{W}} = (1, \ldots, 1, \sqrt{a_0})'$, $\boldsymbol{\mathcal{Z}} = (\boldsymbol{\beta}_1, \ldots, \boldsymbol{\beta}_n, \sqrt{a_0}\boldsymbol{B}_0)'$, and let

$$\begin{aligned} \tilde{\boldsymbol{B}} &= (\mathbf{1}'\mathbf{1} + a_0)^{-1}(\mathbf{1}'\mathbf{1}\bar{\boldsymbol{\beta}} + a_0\boldsymbol{B}_0) \\ &= (n + a_0)^{-1}(n\bar{\boldsymbol{\beta}} + a_0\boldsymbol{B}_0). \end{aligned}$$

Then (B.5) becomes

$$\begin{aligned} P(\boldsymbol{B}, \boldsymbol{\Sigma}_\beta|\cdot) &\propto |\boldsymbol{\Sigma}_\beta|^{-(\nu_0+p+1)/2} exp\left\{-\frac{1}{2}\boldsymbol{V}_0\boldsymbol{\Sigma}_\beta^{-1}\right\} \times \\ &\quad |\boldsymbol{\Sigma}_\beta|^{-(n+p)/2} exp\left\{-\frac{1}{2}(\boldsymbol{Z} - \boldsymbol{W}\tilde{\boldsymbol{B}})'(\boldsymbol{Z} - \boldsymbol{W}\tilde{\boldsymbol{B}})\boldsymbol{\Sigma}_\beta^{-1} + (\boldsymbol{B} - \tilde{\boldsymbol{B}})'(\boldsymbol{B} - \tilde{\boldsymbol{B}})\boldsymbol{\Sigma}_\beta^{-1}\right\} \\ &\propto |\boldsymbol{\Sigma}_\beta|^{-(\nu_0+n+p+1)/2} exp\left\{-\frac{1}{2}\boldsymbol{V}_0\boldsymbol{\Sigma}_\beta^{-1} + \frac{1}{2}(\boldsymbol{Z} - \boldsymbol{W}\tilde{\boldsymbol{B}})'(\boldsymbol{Z} - \boldsymbol{W}\tilde{\boldsymbol{B}})\boldsymbol{\Sigma}_\beta^{-1}\right\} \\ &\quad |\boldsymbol{\Sigma}_\beta|^{1/2} exp\left\{\frac{1}{2}(\boldsymbol{B} - \tilde{\boldsymbol{B}})'(n + a_0)(\boldsymbol{B} - \tilde{\boldsymbol{B}})\}\boldsymbol{\Sigma}_\beta^{-1}\right\} \end{aligned} \tag{B.6}$$

Through the combining of terms, the posterior distribution for $\boldsymbol{B}|\boldsymbol{\Sigma}_\beta$ and $\boldsymbol{\Sigma}_\beta$ is a product of a $\mathcal{MVN}$ and $\mathcal{IW}$, i.e.

$$\begin{aligned} \boldsymbol{\Sigma}_\beta &\sim \mathcal{IW}\left(\nu_0 + n, \boldsymbol{V}_0 + (\boldsymbol{Z} - \boldsymbol{W}\tilde{\boldsymbol{B}})'(\boldsymbol{Z} - \boldsymbol{W}\tilde{\boldsymbol{B}})\right) \\ &\sim \mathcal{IW}\left(\nu_0 + n, \boldsymbol{V}_0 + \left\{(\boldsymbol{\beta}_i - \mathbf{1}\tilde{\boldsymbol{B}})'\ (\boldsymbol{\beta}_i - \mathbf{1}\tilde{\boldsymbol{B}}) + (\tilde{\boldsymbol{B}} - \boldsymbol{B}_0)'\, a_0\, (\tilde{\boldsymbol{B}} - \boldsymbol{B}_0)\right\}\right) \\ \boldsymbol{B}|\boldsymbol{\Sigma}_\beta &\sim \mathcal{MVN}\left(\tilde{\boldsymbol{B}}, \frac{\boldsymbol{\Sigma}_\beta}{(n + a_0)}\right), \end{aligned}$$

where $\tilde{\boldsymbol{B}} = \dfrac{n\bar{\boldsymbol{\beta}} + a_0\boldsymbol{B}_0}{n + a_0}$, and $\bar{\boldsymbol{\beta}} = (\mathbf{1}'\mathbf{1})^{-1}\,\mathbf{1}'\boldsymbol{\beta}$, consistent with the definitions given above.

Appendix C

Glossary for Chapter 4

i	Subscript i identifies components belonging to the i^{th} fur seal
j	Subscript j identifies the j^{th} segment of a fur seal's at-sea track; $j = 1, \ldots, J_i$
J_i	Total number of segments for i^{th} fur seal
$\boldsymbol{y}_i$	The usual multinomial logit response variable belonging to the i^{th} fur seal
$\boldsymbol{W}_i$	Observations related to $\boldsymbol{X}_i$, measured with error
$\boldsymbol{X}_i$	Unobserved "true" covariate(s)
$\boldsymbol{U}_i$	Measurement error in an error model
$\boldsymbol{Z}_i$	Covariate(s) measured without error.
m	Number of regression parameters in lower- and upper-level regression models
n	Number of fur seals (i =1, ..., n=11)
q	Number of higher level covariates in matrix $\boldsymbol{\mathcal{X}}$ ($q = 1$)
$\boldsymbol{\theta}$	the set of unit-level lower-order model unknowns
$\boldsymbol{\psi}$	the set of prior parameters of lower-order parameters ($\boldsymbol{B}, \boldsymbol{\Sigma}$)
$\boldsymbol{\phi}$	the set of hyperparameters in the higher-order model ($\nu_0, a_0, \boldsymbol{V}_0, \boldsymbol{B}_0$)
$\boldsymbol{\Theta}$	the set of all model unknowns; includes $\boldsymbol{\theta}, \boldsymbol{\psi}, \boldsymbol{\phi}$

$p_{ij}^{(1)}$	baseline probability of j^{th} segment of track from i^{th} fur seal being behaviour class "non diving"
$p_{ij}^{(2)}$	probability of being behaviour class "exploratory diving"
$p_{ij}^{(3)}$	probability of being behaviour class "active diving"

Appendix D

MCMC Convergence for Selected Hierarchical Model in Chapter 5

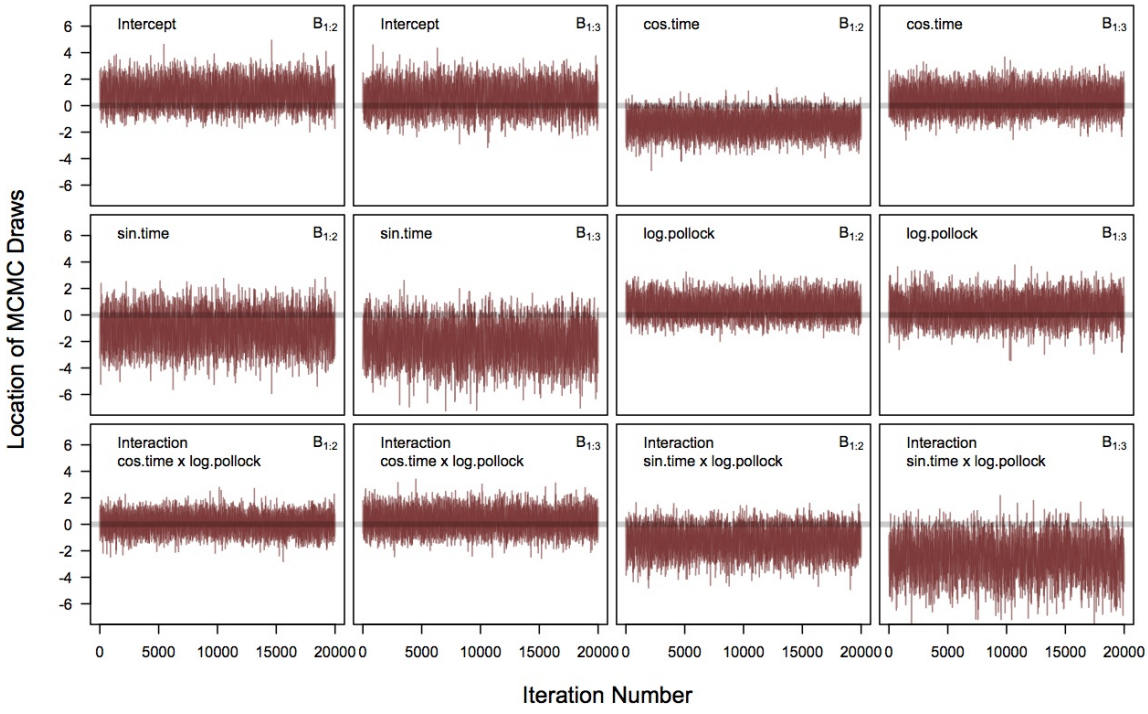

Figure D.1: Thinned 1,000,000 element Gibbs Sampling chain showing draws from the posterior distribution of twelve regression parameters from the population level model $\boldsymbol{B}$ measured with error. Panels show the trace plot for each of the MC Markov chains for population-level parameters linking northern fur seal to at-sea habitat. $B_{1:2}$ and $B_{1:3}$ denote the regression parameters corresponding to $log(p_{ij}^{(2)}/p_{ij}^{(1)})$, and $log(p_{ij}^{(3)}/p_{ij}^{(1)})$, or $(exploring : baseline)$ and $(active : baseline)$ respectively.

Table D.1: Posterior summaries for higher level model coefficients $\boldsymbol{B}$. Similar to Figure (D.1), $B_{1:2}$ and $B_{1:3}$ denote the regression parameters corresponding to $log(p_{ij}^{(2)}/p_{ij}^{(1)})$, and $log(p_{ij}^{(3)}/p_{ij}^{(1)})$ respectively. The column labeled $\hat{R}$ is the Gelman-Rubin Bayesian measure of convergence.

				Median	Credible Interval		
		Mean	St. Dev.	$q_{.5}$	$q_{.025}$	$q_{.975}$	$\hat{R}$
Intercept	$\boldsymbol{B}_{1:2}$	0.932	0.772	0.928	-0.564	2.473	1.01
	$\boldsymbol{B}_{1:3}$	0.645	0.828	0.649	-0.983	2.277	1.00
cos(time)	$\boldsymbol{B}_{1:2}$	-1.352	0.613	-1.336	-2.594	-0.179	1.02
	$\boldsymbol{B}_{1:3}$	0.422	0.680	0.413	-0.874	1.797	1.00
sin(time)	$\boldsymbol{B}_{1:2}$	-1.144	1.055	-1.146	-3.216	0.913	1.00
	$\boldsymbol{B}_{1:3}$	-2.111	1.175	-2.098	-4.434	0.159	1.04
log(pollock)	$\boldsymbol{B}_{1:2}$	0.622	0.633	0.610	-0.597	1.897	1.03
	$\boldsymbol{B}_{1:3}$	0.477	0.762	0.478	-1.011	1.974	1.03
cos(time) × log(pollock)	$\boldsymbol{B}_{1:2}$	0.064	0.549	0.068	-1.045	1.131	1.01
	$\boldsymbol{B}_{1:3}$	0.264	0.645	0.257	-1.005	1.549	1.01
sin(time) × log(pollock)	$\boldsymbol{B}_{1:2}$	-1.258	0.789	-1.239	-2.866	0.252	1.00
	$\boldsymbol{B}_{1:3}$	-2.479	1.177	-2.419	-4.993	-0.339	1.05

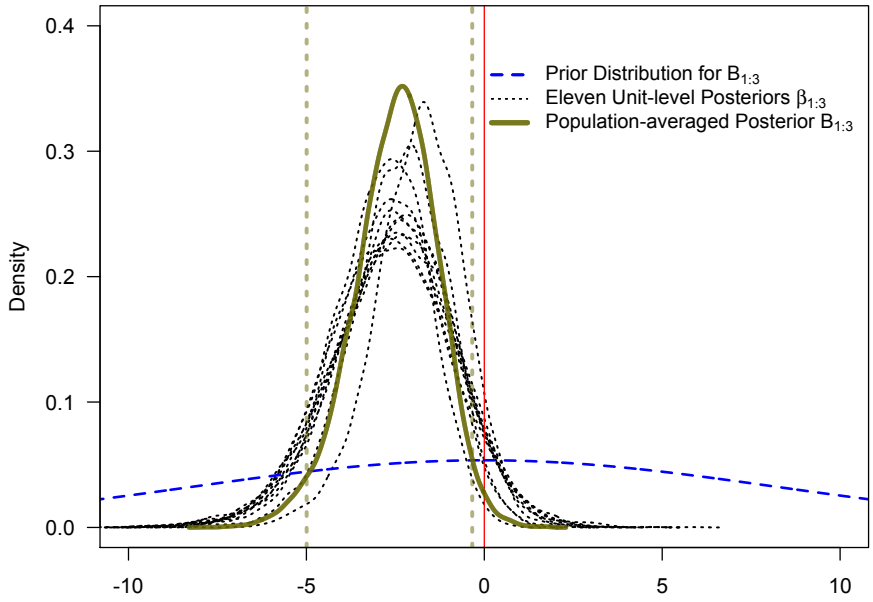

Figure D.2: Comparison of prior distribution and unit- (lower-) level and population-averaged (upper-level) posterior distributions of sin.time × log.pollock $\boldsymbol{B}_{1:3}$ in final Time × Pollock model. In this hierarchical framework, the posterior of $\boldsymbol{B}_{1:3}$ becomes part of the prior distribution for the equivalent eleven $\boldsymbol{\beta}_{1:3}$ unit-level parameters. The vertical dashed 95% Credible Interval for $\boldsymbol{B}_{1:3}$ does not include 0 (vertical red line). The unit-level estimators of $\boldsymbol{\beta}_{1:3}$ undergo shrinkage, and will exhibit less variation that least squares estimates computers equation by equation. The amount of shrinkage will be dictated both by the prior hyper-parameters and the data (Gelman et al. 2004).

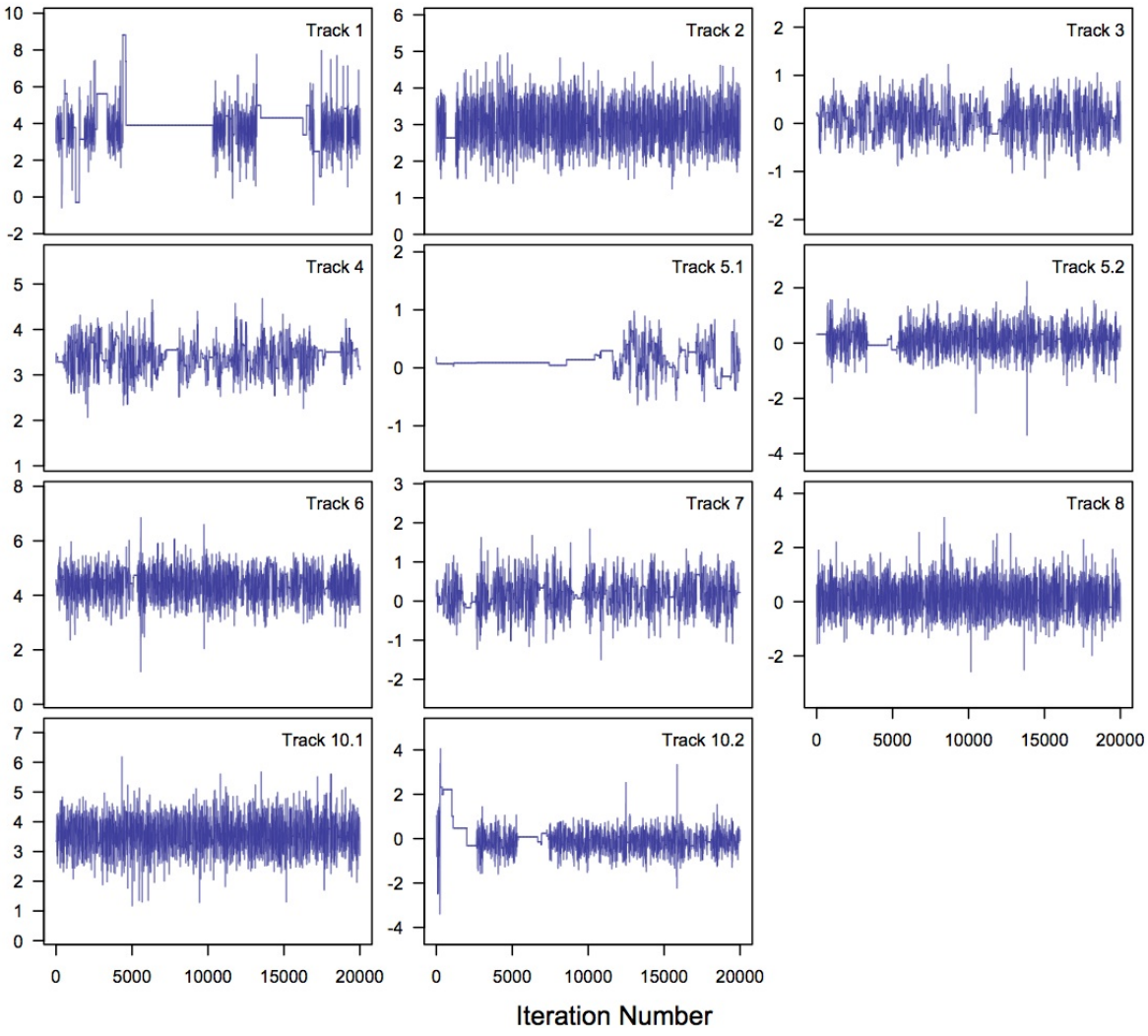

Figure D.3: An example of a thinned 1,000,000 element Metropolis Hastings chain showing draws from the posterior distribution of a covariate $\boldsymbol{X}$ measured with error. Panels show the trace plot for each of the eleven female northern fur seal tracks and the covariate pollock catch modelled with Berkson error model.

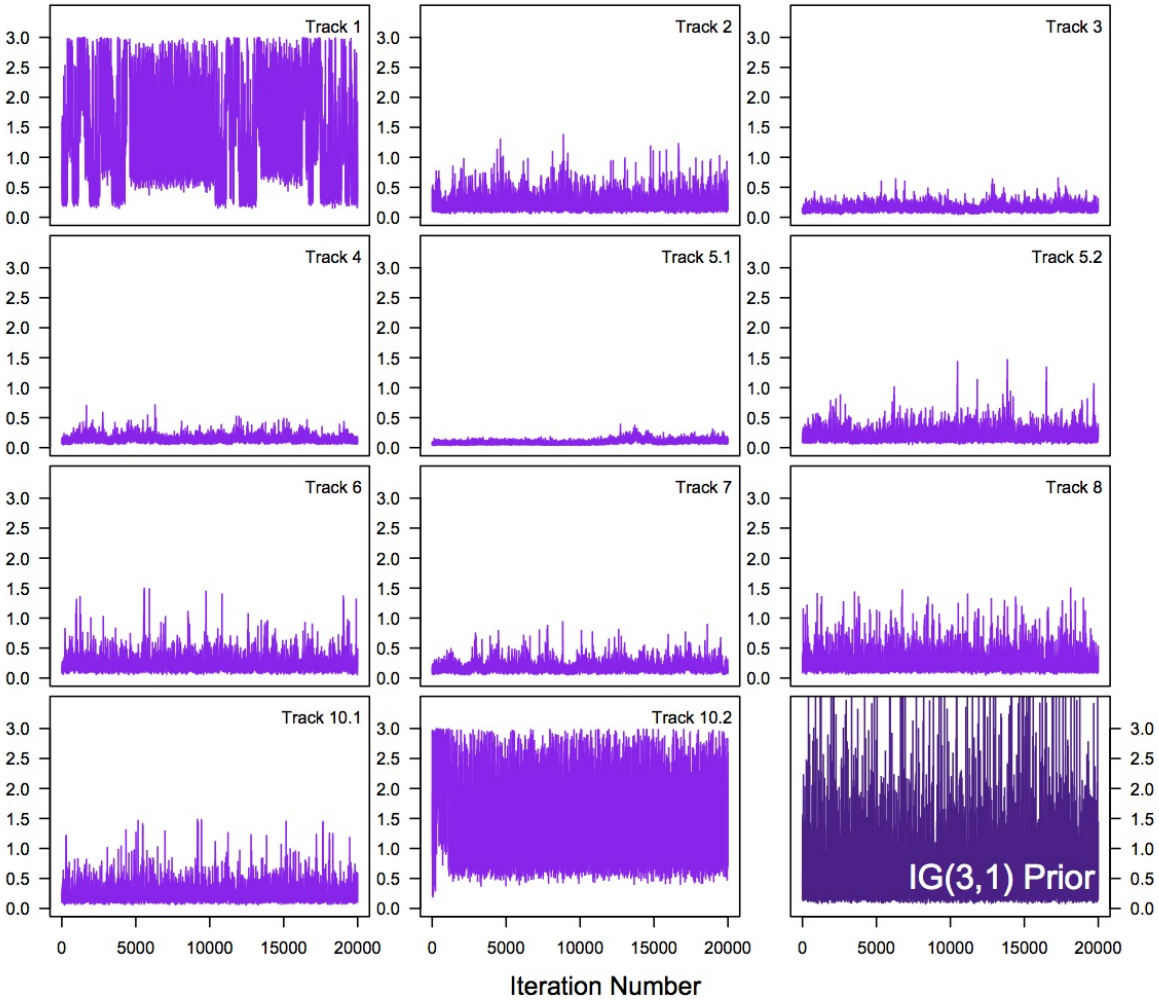

Figure D.4: An example of a thinned 1,000,000 element Metropolis Hastings chain showing draws from the conjugate posterior $\mathcal{IG}$ distribution that describes the measurement error in haul size of pollock (σ^2_{iU}). Panels show the trace plot for each of the eleven female northern fur seal tracks and the measurement error around pollock catch modelled with a Berkson error model. Any posterior samples that were outside of five standard deviations of the mean were rejected. This affected 2 of the 11 lower level models, Tracks 1 and 10.2. The final panel shows the prior distribution assigned to σ_{iU}.