STATISTICAL MODELLING OF MARINE FISH POPULATIONS AND COMMUNITIES

GARABET PANIKIAN
Doctor of Philosophy

University of York
Computer Science

September 2016

Glory and praise to our God the Father almighty through whom everything was made and through His only Son our Lord Jesus Christ, conceived by the power of the Holy Spirit, we were delivered from slavery and brought to light: many are His blessings, His mercy is infinite and His kingdom shall have no end.

Abstract

Sustainable fisheries management require an understanding of the relationship between the adult population and the number of juveniles successfully added to that population each year. The process driving larval survival to enter a given stage of a fish population is highly variable and this pattern of variability reflects the strength of density-dependent mortality. Marine ecosystems are generally threatened by climate change and overfishing; the coupling of these two sources have encouraged scientists to develop end-to-end ecosystem models to study the interactions of organisms at different trophic levels and to understand their behaviours in response to climate change. Our understanding of this important and massively complex system has been constrained historically by the limited amount of data available. Recent technological advances are beginning to address this lack of data, but there is an urgent need for careful statistical methodology to synthesise this information and to make reliable predictions based upon it.

In this thesis I developed methodologies specifically designed to interpret the patterns of variability in recruitment by accurately estimating the degree of heteroscedasticity in 90 published stock-recruitment datasets. To better estimate the accuracy of model parameters, I employed a Bayesian hierarchical modelling framework and applied this to multiple sets of fish populations with different model structures. Finally, I developed an end-to-end ecological model that takes into account biotic and abiotic factors, together with data on the fish communities, to assess the organisation of the marine ecosystem and to investigate the potential effects of weather or climate changes.

The work developed within this thesis highlights the importance of statistical methods in estimating the patterns of variability and community structure in fish populations as well as describing the way organisms and environmental factors interact within an ecosystem.

Contents

List of Tables

List of Figures

List of Algorithms

Acknowledgements

Sincere thanks to my supervisor Dr. James Cussens for all his help and encouragement, and for being such an excellent supervisor. Thanks for the numerous meetings (111 meetings on an average of 80 minutes per meeting) that you provided me and gave me the freedom to explore different areas so as to increase my knowledge.

Sincere thanks to my assessor Dr. Jonathan Pitchford for bringing to my attention the ecological stock-recruitment time series datasets; for the useful discussions, directions and motivations, and time spent for reviewing and polishing my work. Thank you to Dr. Suresh Manandhar and Dr. Jon Barry (Cefas) for a very useful viva and for providing me with a very thorough feedback.

Special thanks to Dr. Richard Everitt (University of Reading, UK) who advised me in the early phase of my research to focus on selecting an appropriate dataset first and then to build my research.

Special thanks to my friend Dr. Peter Whitehead for his support, remarks, and long discussions that he provided me during my journey. Many thanks to Dr. Kashif Khan for the useful discussions that he provided me during the early phase of my research.

I especially acknowledge Dr. Daniel Ricard (Fisheries and Oceans Canada) for helping me in understanding the VPA type assessment of stock-recruitment populations and share his opinion over several topics of my research. I acknowledge Professor Bradley Efron (Stanford University, USA) for his guidance concerning the parametric bootstrap method. I would like to acknowledge Dr. Gustav Delius who helped me in identifying the community based model of fish populations.

I especially acknowledge Dr. David Johns (SAHFOS, UK) for his support and helpful feedback on Chapter 5 and for the planktonic data that he provided me.

During my research, I attended four Graduate Training Programme (GTP) summer school courses in the Newcastle University:

- GTP 2012, Advanced Computational Bayesian Inference.
- GTP 2013, Highly Structured Stochastic Systems.
- GTP 2014, Analysis of Spatial and Temporal data.
- GTP 2014, Bayesian Modelling for Systems Biology.

These courses have enriched my statistical knowledge and helped me to understand and implement various MCMC algorithms.
Many thanks to Dr. Kokouvi Gamado (BioSS, UK) who helped me in implementing the Bayesian hierarchical model in JAGS. Many thanks to Sarah Christmas who helped me in many administrative tasks during the course of my research.
I should also thank my parent for their love and encouragement during my life; thank to my parent-in-law who encouraged me in completing my degree.
I dedicate this research to my young children: Ilaria, Clara and Anthony and most of all to my beloved wife Myrna for her endless support, love and for being with me on every step of this journey.
Finally, I praise my Lord Jesus Christ for giving me wisdom and strength to complete this degree successfully.

Declaration

I declare that this thesis is a presentation of original work and I am the sole author. All sources are acknowledged as References. No portion of the work referred to in this thesis has been submitted in support of an application for another degree or qualification of this or any other university or other institute of learning. Parts of this thesis have been published or are in submission:

- Panikian, G. and Cussens, J. and Pitchford, W.J. (2015): Identification and quantification of heteroscedasticity in stock-recruitment relationships. *Canadian Journal of Fisheries and Aquatic Sciences*, NRC Research Press, 72, pp 1259-1271, 2015.

- Panikian, G. and Cussens, J. and Pitchford, W.J. (2017): Bayesian Hierarchical Modelling to identify community-level processes: the Stock-Recruitment relationship on Georges Bank. *In submission.*

- Panikian, G. and Cussens, J. and Pitchford, W.J. (2017): End-To-End Statistical Modelling for Marine Ecosystems via Machine Learning. *In preparation.*

Chapter 1

Introduction

Every truth without exception —and whoever
may utter it —is from the Holy Spirit.
–St. Thomas Aquinas.

In this chapter I present an overview of my work and define the problem that I am trying to solve using statistical analysis methods. I describe growth and mortality rates as the driving forces for provoking the dynamical behaviour of fish populations that depend on environmental changes and fishing activities. Several methods were used to analyse fish population dynamics. Among those I provide an overview of some methods that I believe are in line with my perspective. Finally, I conclude this chapter with an outline of the structure of the remainder of the report.

1.1 Motivation and Overview

The oceans cover 71% of the earth's surface and represent a habitat for many species of fish and sea life. marine systems contribute 27% to global carbon budgets (Le Quéré et al., 2015), and more than 3.1 billion people (20% of the world's population) depend on fish as their main source of protein (Food and Agriculture Organisation, 2016). The Food and Agriculture Organisation (2016) of the United Nations stated that around 90% of the world's stocks are either fully fished or overfished and the total world fishery production (capture plus aquaculture) is projected to increase by 31 million tonnes in the next decade to reach 178 million tonnes in 2025, which is an increase of 17% to the global consumption of fish

supplies. Unfortunately, fish species are limited in quantities and the use of illegal and inappropriate fishing activities could lead to their extinction. To protect these resources, proper rules and regulations in fisheries management should be implemented; local policies should promote aquaculture as an alternative form for animal food production, which involves cultivating freshwater and saltwater populations under controlled conditions. The United Nations Food and Agriculture Organisation states that freshwater species such as: carp, catfish and tilapia, will account for most of the increase in aquaculture production and represent about 60% of total aquaculture production in 2025; however, production of higher-value species, such as shrimps, salmon and trout, is also projected to continue to grow in the next decade.

In this research, I exploit a set of statistical methods to reveal information embedded in fish stock assessments. I analysed whether adding an extra parameter (non-constant variance) η_1 to the stock-recruitment (S-R) relationship can explain better the variability in fish recruitment. The analysis is based on two different approaches: the first is based on using both frequentist and Bayesian paradigms applied to single fish stock assessment; and the second consists of employing a Bayesian hierarchical framework applied on multiple fish populations with different model structures. Finally, I invented a new end-to-end model to analyse the impact of climate change, variability and extreme weather events on the abundance of marine species.

1.2 Fish population dynamics

The dynamical behaviour of a fish population is the way a population of fish varies over time. This variability is caused by the growth and mortality rates in a population. Hilborn and Walters (1992) define a fish population as a unit stock of a homogeneous collection of fish that are all subject to the same opportunities for growth and reproduction and the same risks of natural and fishing mortality. A sufficient amount of nutrition is necessary for fish to gain weight and reproduce as they age. Steele et al. (1977, page 44) proposed a relationship between food intake and body weight or body length, where they illustrated how the growth is sensitive to fluctuations in the consumption of food. Obvious sources of mortality are as follows: fishing activities; emigration interpreted as mortality; immigration interpreted as negative mortality; predation; starvation; poisonous pollutants and

so forth (Steele et al., 1977, page 87). The mortality is often partitioned into three different categories: (i) fishing mortality, (ii) density-dependent mortality, and (iii) density-independent mortality. The density-dependent mortality rate of fish populations rises when the carying capacity becomes saturated or when food resources diminish; however, the density-independent mortality rate rises when the environmental conditions become harsh —chemical changes (salinity content, oxygen concentration, and acidification) of the aquatic environment. After hatching the individuals in a cohort are subject to an exponential decay over time, meaning that the higher the mortality rate the faster the population numbers decline. Therefore, we can write the population size of next year as a survival rate s times population size of this year plus new recruits (Hilborn and Walters, 1992), such that

$$\begin{aligned} N_{t+1} &= sN_t + R_{t+1} \\ &= N_t \exp\{-f - (m_1 + m_2 + \ldots + m_n)\} + R_{t+1} \\ &= N_t \exp(-f - m) + R_{t+1}, \end{aligned} \tag{1.1}$$

where f is the fishing mortality rate and $m = (m_1 + m_2 + \ldots + m_n)$ is the total natural mortality that includes both density-dependent and density-independent factors.

1.3 Model of stock and recruitment

The Spawning Stock Biomass constitutes a fundamental characteristic in fisheries science to assess the growth of a fish population. Recruitment is the abundance in numbers or biomass of juvenile fish that live from hatching to adult life. The ideal assessment of spawning stock is the number of eggs produced by adult fish; marine biologists often assess the spawing stock as the total weight of the fish in a stock capable of reproducing, which is usually measured in terms of biomass (e.g. tonnes) (Shepherd and Cushing, 1990). Needle (2001) reviewed a synopsis of the types of recruitment model that are utilised in stock assessments and the degree to which these models are employed; he also emphasised the need to link biological and oceanographic recruitment models with assessment procedures to study the problem of fish stock assessment. In this context, I briefly describe the most commonly used models: Beverton-Holt, Ricker and Deriso-Schnute models.

The simplest model that relates recruitment to spawning stock is the Beverton and Holt (1957) model, which can be described as

$$R = \frac{\alpha S}{\beta + S}, \tag{1.2}$$

where R is the recruitment, S is the stock biomass, α measures the productivity and β represents the density-dependent mortality in a population. As the stock size gets very large, the recruitment value tends asymptotically to α.
An alternative formulation of the Beverton-Holt model incorporates a parameter characterising the 'steepness' of the stock-recruit relationship at low stock sizes. Mace and Doonan (1988) defined the steepness (h) as the fraction of recruitment from an unfished population (R_0) when the spawning stock biomass (SSB) is at 20% of its unfished level (S_0). As h approaches 1, the Beverton-Holt relationship approaches a form in which recruitment is constant; when h is 0.2, recruitment is linearly related to SSB. The advantage of this formulation is that h is unaffected by the actual size of the stock.
Ricker (1954) suggested an exponential decay of recruitment as the spawning stock biomass increases, such that

$$R = \alpha S \exp(-\beta S). \tag{1.3}$$

As $S \to R$, then $R = \log(\alpha)/\beta$.
Deriso (1980) introduced the general three-parameter stock-recruitment relationship to incorporate both Ricker and Beverton-Holt models, which was further developed by (Schnute, 1985): resulting in the so-called the Deriso-Schnute model:

$$R = \alpha S(1 - \gamma\beta S)^{1/\gamma}, \tag{1.4}$$

where the parameters α and β measure the productivity and the density-dependent mortality in a population; however, γ enables us to choose between different survival models, such that

$$R = \begin{cases} \alpha S, & \lim \gamma \to -\infty \\ \dfrac{\alpha S}{1+\beta S}, & \gamma = -1 \\ \alpha S \exp(-\beta S), & \lim \gamma \to 0 \\ \alpha S(1-\beta S), & \gamma = +1. \end{cases} \tag{1.5}$$

If $\gamma \to -\infty$ the stock-recruitment relationship becomes linear as the density-dependent effect weakens; however, if γ becomes positive such as $\gamma = +1$, the density-dependent mortality rate increases and fish recruitment becomes attenuated with large stock size. Fish populations pass through a number of life-history stages, starting from planktonic egg to larval to juvenile, before recruitment and then adult stages. The density dependent mortality is believed to influence fish populations most during the juvenile stage (Myers and Cadigan, 1993a); Hjort (1914) considered that recruitment variability was determined in the early stages of larval development. Much of the research on the early life history of fish has focused on Hjort's starvation hypothesis which implies a high mortality of small feeding larvae during the first few weeks of life (Wooster and Bailey, 1989).

1.4 Methods used to analyse fish populations

Several models have been proposed to describe how a fish population changes over time. The exponential law of population growth (Malthus, 1798) is probably the most commonly used method in the field of population ecology to represent dynamic populations, which is regarded as the first principle of population dynamics (Turchin, 2001). Conservation laws in modelling a population assume that birth, death, emigration, and immigration are proportional to the number of organisms in the population (Turchin, 2001). On the other hand, Bayesian statistical methods have also been investigated by many scientists for enhancing fishery management systems. For example, McAllister and Kirkwood (1998) compared the performance of two Bayesian models for fitting a logistic model to relative abundance of fish species: the first employs a non-conjugate reference prior that uses all historic data; the second employs a conjugate prior that

provides closed form analytical solutions. Munch et al. (2005) used a Bayesian non-parametric approach with conjugate prior knowledge to evaluate uncertainties in fishery management systems (e.g. stock-recruitment curve, steepness and the stock biomass at maximum sustainable yield). Their method was applied to different synthetic data sets generated from a variety of parametric models (e.g. Ricker, Beverton-Holt, Shepherd, SailaLorda and Open-mixture), and they finally tested it on empirical datasets for lingcod *Ophiodon elongatus* and several salmonids as they found a comparable fit to the Ricker and Beverton-Holt models. Brodziak and Piner (2010) analysed the North Pacific striped marlin because they found it vulnerable to recruitment overfishing in pelagic longline fisheries which targeting tunas. The authors applied two different scenarios to account for different hypotheses about the steepness of the stock-recruitment relationship ($h = 0.7$ and $h = 1$) for which Beverton-Holt and Ricker models were applied to estimate the maximum sustainable yield (S_{MSY}) and the associated limit fishing mortality (F_{MSY}) respectively. Results were then combined by the mean of model averaging to assess the probable status of a fishery resource for these competing assessment scenarios.

Bayesian hierarchical analysis of stock-recruitment relationship offers a natural way to incorporate multiple stock assessment hypotheses. Michielsens and McAllister (2004) applied a Bayesian hierarchical analysis using the Beverton-Holt and Ricker models over the Atlantic salmon stock-recruitment data to infer the steepness parameter for Baltic salmon for which no data are observed. What makes the steepness parameter interesting is its characteristic that is comparable among populations.

Chen and Fournier (1999) proposed a Bayesian inference method based on a mixture distribution function to simulate a heavy-tailed distribution for analysing fisheries data contaminated with outliers. This has the effect of preventing the probability of occurrence of an event to drop off quickly as we move away from the centre of the distribution. Moreover, state space models were applied for modelling the dynamical behaviour of discrete-time population. Buckland et al. (2004) modelled the wildlife population process by the state process and measurements by the observation processes. The evolution from one state to another is described by three separate first-order Markov sub-processes (i.e. survival, movement and birth), which are linked together to produce the overall population dynamics model. These might correspond to winter survival, movement in

spring, and births in early summer. Sequential Monte Carlo was used for evaluating model parameters as it avoids computing the intractable likelihood. The authors proposed an interesting extension to their work by applying time-varying parameters in a hierarchical framework, as in Newman (2000).
Finally, Tsitsika et al. (2007) applied univariate and multivariate autoregressive integrated moving average models to predict the total pelagic fish production time series data of monthly catches up to 12 months in advance. The univariate model was constructed as a linear function of past values of the time series; however, the multivariate model was developed so as to predict the value of one species including the effect of other species. Apparently, the fitting accuracy of multivariate models outperformed the univariate ones as they incorporated additional information in the model.

1.5 Aims and Objectives

The principal aim of this thesis is to develop a statistical inference method for dynamical fish populations following a discrete time system. A key feature of this research is to understand the factors that are controlling this evolution, and come up with a model capable of capturing the underlying structure of the time-series data. As application I choose to study the influential factors that affect recruitment variability of fish population in oceans, and develop an end-to-end statistical model using biotic, abiotic and fish populations to assess the impact of weather change on the marine ecosystem.
To achieve this aim, it is necessary to develop some research objectives such as to:

1. Review literature about existing ecological systems (e.g. linear and non-linear models) to understand how one can mathematically represent the dynamical behaviour of fish populations.

2. Review literature about the relationship between survival variability and the strength of density dependence to understand the characteristics of population regulation.

3. Apply a statistical analysis using Bayesian and frequentist paradigms to assess the reliability of a non-constant variance added to the Deriso-Schnute model.

4. Apply Bayesian hierarchical models to combine knowledge of fish stocks so as to estimate the distribution of unobserved parameters and understand the best model structure that can explain the variability of fish recruitment.

5. Develop an end-to-end model to understand the impact of climate change on the ecosystem and marine species.

1.6 Problem statements

The problem statements of my doctoral research was to investigate the following questions:

- Can heteroscedastic (non-constant variance) models explain why there is a high survival rate in fish populations at low stock size?

- Can Bayesian hierarchical analysis improve the estimation of key parameters and achieve a better fish recruitment prediction for harvested fish populations?

- Can we predict the impact of weather change on the ecological system using end-to-end ecological models?

The answer for each of these questions is described in Chapters: 3, 4 and 5 respectively.

1.7 Data Used

In this research I use the time series for the spawning stock dataset taken from two sources: (i) RAM Legacy Stock Assessment Database (Ricard et al., 2012), and (ii) International Council for the Exploration of the Sea (ICES) (`http://standardgraphs.ices.dk/stockList.aspx`). The planktonic data are taken from the Sir Alister Hardy Foundation for Ocean Science (SAHFOS) (Johns, 2015), and the abiotic variables are taken from National Oceanic and Atmospheric Administration (NOAA) (NOAA, 2015).
In this work, I divided each dataset into three disjoint subsets composed of: training, validation and testing sets. The training set consists of training the models, the validation set consists of validating and tweeking the parameters of models

and the testing set consists of testing the models. A common splitting choice in machine learning is to choose the first 60% for the training, the next 20% for validation and the remaining 20% for testing, and this I chose to do in my whole thesis.

1.8 Structure of the Report

The remainder of this report is organised as follows:

Chapter 2 is the literature review of this thesis for which relevant statistical information is being explored and analysed. I begin by describing the theory of probability, compare frequentist and Bayesian approaches, and review approximation techniques such as: functional approximations and Markov chains. Other methods are reviewed in this chapter such as: Bayesian hierarchical models, dynamic Bayesian networks, non-parametric models, and time series processes that might evolve over time.

Chapter 3 consists of presenting an analysis on assessing the reliability of the non-constant variance (heteroscedasticity) in the stock-recruitment models using both frequentist and Bayesian methods. The non-constant variance is applied to a global compilation of stock and recruitment data to examine its influence on the relationship between the variability in survival and population abundance. Much of the work in this chapter is published in the *Canadian Journal of Fisheries and Aquatic Sciences*, printed in Appendix H.

Chapter 4 consists of presenting different hierarchical Bayesian models for improving estimation of key parameters found in stock-recruitment relationships (i.e. the non-constant variance) and improving our understanding of dynamical behaviour of fish populations and community structures.

Chapter 5 consists of presenting a simple end-to-end modelling framework exploiting theory from dynamic Bayesian networks for coupling environmental, planktonic and fisheries data to arrive at predictive ecosystem-scale models.

Chapter 6 consists of summarising the principal findings and contributions of my research and to identify possible avenues for future work.

Chapter 2

Background

In this chapter, I review the necessary literature in statistical learning to represent an appropriate background framework for this research. In particular I focus on both frequentist and Bayesian inferential theories for learning models from data. Additionally, I review basic principles of dynamical Bayesian networks, approximate inference methods, advantages of non-parametric models and time series processes that might develop over time.

2.1 Basic Concepts in Probability Theory

This section consists of describing basic rules and elementary concepts in probability theory.

2.1.1 Uncertainty

Uncertainty is a term used to express our ignorance about events or about measurements that we have already performed. It can be caused by the lack of sufficient information, knowledge or precision. This principle is applicable to a wide range of applications and fields such as: politics, economics, physics, engineering, biology, law and so forth. Among the numerous practical cases, the following examples illustrate how uncertainty arises from one or multiple sources and impacts on our ability to make decisions.
Lack of information as well as having limited knowledge in understanding real problems may render us unsure about our judgments. For instance, the jury would become uncertain about convicting the offender if they could not find real

evidence that proved his guilt. On the other hand, an inappropriate software development methodology can lead to inaccurate or invalid results and hence throw the project behind schedule. In medical diagnosis, the presence of some symptoms are not always enough to diagnose the disease with a high degree of certainty. Though, additional medical tests are usually required to be performed (Castillo et al., 1997, page 4). From a scientific point of view, uncertainty constitutes an integral part of all sciences, which is a fundamental measure that characterises the degree of confidence when measuring a physical phenomenon such as: measuring temperature, length, voltage and so forth. Heisenberg (1927) conducted an experiment to measure simultaneously the position and momentum of an electron particle, when struck by a photon. Accordingly, he concluded his work with the following statement:

> *The more precisely the position is determined, the less precisely the momentum is known in this instant, and vice versa.*

This was a significant step forward, for that time period, in the development of the modern theory of quantum mechanics. An intuitive measure of uncertainty is *probability*, which we are going to illustrate subsequently.

2.1.2 Probability theory

Early civilisations like the Egyptians, Babylonians and Greeks theoreticians laid down basic geometry and algebra techniques but they had not come across chance. This has been on hold until the renaissance (17^{th} century) when Chevalier de Meré asked Pascal about how to figure out, at any given stage, the probability of winning a gambling game. Then Pascal involved Fermat on the subject of this question (Todhunter, 1865, page 7). In these early days, the theory of probability was based either on the relative frequency of occurrence of an event or on the subjective degree of belief a person has in an event. Let us assume that we may repeat an experiment a large number of times, and we query the system about the relative frequency of occurrence of an event e. If we denote by n the number of repetitions that are assumed to be independent and identically distributed (i.i.d.) and by S_n the number of times the event e was found to be true, we could then state if n is very large the ratio S_n/n should be near to the probability p of the event e. To make a precise mathematical formulation of this statement, we translate the term i.i.d. as 'Bernoulli trials' with probability p for success, then

the average number of successes S_n/n should converge to p. Chebichef (1846) proved this hypothesis by assuming the probability that S_n/n exceeds $p+\epsilon$, where $\epsilon > 0$, in the form

$$\Pr\left(\frac{S_n}{n} > p+\epsilon\right) = \Pr\left(\frac{S_n}{n} - p > \epsilon\right).$$

An upper bound of this probability is given by Chebyshev's inequality, such that

$$\Pr\left(\left|\frac{S_n}{n} - p\right| \geq \epsilon\right) \leq \frac{\sigma^2}{n\epsilon^2}, \quad \text{where } \sigma^2 \text{ is the variance parameter.}$$

$$\Rightarrow \Pr\left(\left|\frac{S_n}{n} - p\right| < \epsilon\right) = 1 - \Pr\left(\left|\frac{S_n}{n} - p\right| \geq \epsilon\right) \geq 1 - \frac{\sigma^2}{n\epsilon^2}.$$

Therefore, we conclude that in the limit of large number of trials the probability of average number of successes deviates from p by a small number ϵ, such that

$$\lim_{n\to\infty} \Pr\left\{\left|\frac{S_n}{n} - p\right| < \epsilon\right\} \to 1. \tag{2.1}$$

This theory has been improved by many other scientists such as de-Moivre, Laplace and in addition to so many others; but it had not found a precise mathematical formulation for nearly three centuries till Kolmogorov (1933) when he finally imposed new axioms for modern probability theory, defined by

> A triple $(\Omega, \mathscr{U}, \Pr)$ is called a *probability space* if it comprises a set Ω of events, a σ-algebra $\mathscr{U}$ of subsets of Ω, and a probability measure Pr on the pair $(\Omega, \mathscr{U})$, which can be defined as a function Pr: $\mathscr{U} \to [0,1]$ satisfying
>
> 1. The probability of an event is a non-negative real number: $\Pr(e) \geqslant 0$.
> 2. The probability of an event to occur in the entire sample space is one: $\Pr(\Omega) = 1$, and the probability of an event to occur on the empty space is zero: $\Pr(\emptyset) = 0$.
> 3. Let $e_1, \ldots, e_n, \ldots$ be a set of disjoint events of $\mathscr{U}$, in that $e_i \cap e_j = \emptyset$ for all pairs i, j satisfying $i \neq j$, then
>
> $$\Pr\left(\bigcup_{i=1}^{\infty} e_i\right) = \sum_{i=1}^{\infty} \Pr(e_i).$$

This contribution was significantly important for enabling further probability rules to emerge.

2.1.3 Conditional probability

Information sometimes arrives in stages, and this may happen when we are observing a physical process. For example, if we are observing the activity of a hot spring or a geyser bursting hot water skyward, we notice that the eruptions might be occurring either at regular intervals of time or not. While observing these eruptions, one might compute the probability of an event e_2 to occur given an observed event e_1 of the form

$$\Pr(e_2|e_1) = \frac{\Pr(e_1 \cap e_2)}{\Pr(e_1)} \tag{2.2}$$

where e_1 is the event of the eruption at time t_1 and e_2 is the expected event at time t_2. If these events are independent one may write $\Pr(e_1 \cap e_2) = \Pr(e_1)\Pr(e_2)$. More generally, a family $\{E_i : i \in I\}$ is called independent if

$$\Pr\left(\bigcap_{i\in J} E_i\right) = \prod_{i\in J} \Pr(E_i)$$

for all finite subsets J of I.

2.1.4 Bayes theorem

Bayes theorem was invented by Thomas Bayes (1702-1761), which relates the conditional probabilities to their inverses. This approach was used as a measure to represent uncertainty in decisions, known as the first level of inference. Keynes (1921) proposed that the theory of probability is logical, because it involves logical relations between the propositions that express our direct knowledge and the propositions about which we seek indirect knowledge. Soon after, Cox (1946) used these rules as axioms and developed an elegant probability theory using Boolean algebra and degree of belief.
Since conjunction is a commutative operation in Boolean algebra, the conditional probability expressed in Equation (2.2) can be written as

$$\Pr(e_1 \cap e_2) = \Pr(e_2|e_1)\Pr(e_1) = \Pr(e_1|e_2)\Pr(e_2). \tag{2.3}$$

By rearranging we get Bayes' theorem which takes the form

$$\Pr(e_2|e_1) = \frac{\Pr(e_1|e_2)\Pr(e_2)}{\Pr(e_1)}, \tag{2.4}$$

where $\Pr(e_2)$ could represent the prior knowledge which reflects our belief about e_2 before we observe any evidence, $\Pr(e_1)$ is the probability of observing the evidence, $\Pr(e_1|e_2)$ is the likelihood which can be determined from the case histories of the event e_2, and $\Pr(e_2|e_1)$ is the posterior probability that determines our new belief about e_2 after observing the evidence e_1. In other words, the Bayes theorem relates the conditional probabilities of events before and after observing the evidence. Equivalently, one may write Bayes' theorem in words

$$P\text{osterior} = \frac{L\text{ikelihood} \times P\text{rior}}{E\text{vidence}}. \tag{2.5}$$

Bayes' theorem may be highly subjective because of the choice of prior distribution depends upon the individual (or group) concerned. One might develop further the expression of the posterior such that

$$\Pr(e_2|e_1) = \frac{\Pr(e_1|e_2)\Pr(e_2)}{\Pr(e_1|e_2)\Pr(e_2) + \Pr(e_1|\neg e_2)\Pr(\neg e_2)}, \tag{2.6}$$

where $\neg e_2$ is the negation form of e_2, which means that the proposition we made about e_2 is false. Bayes theorem can be used in many applications of probabilistic modelling and inference such as: spam filtering and human motion modelling.

2.1.5 Random variables

A random variable is a function of a sample space of possible numerical values, which can be defined as

> Let $(\Omega, \mathscr{U}, \Pr)$ be a probability space. A mapping
>
> $$X : \Omega \to \mathbb{R}$$
>
> with the property that $\{e \in \Omega : X(e) \leq x\} \in \mathscr{U}$ for each $x \in \mathbb{R}$. We equivalently say that X is $\mathscr{U}$*-measurable.*

For example, the sample space for a fair coin tossed twice (i.e. $x = 2$) consists of the following set $\Omega = \{\text{HH,HT,TH,TT}\}$. For an event $e \in \Omega$, the number of

observed tails are mapped as: $X(e) \leq x$, where $x \in \mathbb{R}$. This can be illustrated in the form

$$\mathrm{X(HH)} = 0,\ \mathrm{X(HT)} = \mathrm{X(TH)} = 1,\ \mathrm{X(TT)} = 2.$$

The values of a random variable X can be either real, discrete or complex. The probability that X will take the value x and Y will take the value y can be written as $\Pr(X = x, Y = y)$, which is called the joint probability of $X = x$ and $Y = y$. For discrete random variables, the two fundamental rules of probability theory can be written in the form

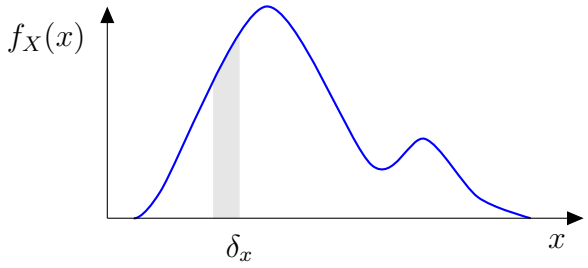
$$\begin{cases} \textbf{Marginal probability} & \Pr(X = x) & = \sum_y \Pr(X = x, Y = y), \\ \textbf{Joint probability} & \Pr(X = x, Y = y) & = \Pr(Y = y | X = x) \Pr(X = x). \end{cases} \tag{2.7}$$

This is also known as the sum rule and the product rule of probability. Throughout this thesis, I am going to use an uppercase letter to represent random variables and a lowercase letter to represent realisations.

2.1.6 Density functions

A continuous random variable X defined over an interval of values is constrained to take its value from within this interval. For example, if X lies in an interval $[a, b]$, the behaviour of X can then be described by a probability density function $f_X(x)$ for $a \leq x \leq b$. Figure 2.1 illustrates the area under the density function

Figure 2.1: Example of a probability density function.



between two points x and $x + \delta_x$. However, the mathematical representation for the distribution function between these points can be defined by

$$\Pr(x < X \leq x + \delta_x) = \int_x^{x+\delta_x} f_X(u)du. \tag{2.8}$$

This can be extended to any measurable set, such that

$$\Pr(X \in A) = \int_A f_X(u)du. \tag{2.9}$$

Additionally, the total area under the density function must be equal to one, so that

$$\int_a^b f_X(x)dx = 1. \tag{2.10}$$

On the other hand, for a discrete random variable the above probability density function turns out to become a probability *mass* function instead, and the integration term will be replaced by summation, which can be used to describe the concentration of possible observed values along the x axis.

2.1.7 Expected values and Covariances

The general expression for the expected value of any function g of X can be written as follows:

$$\mathbb{E}[g(X)] = \int_{-\infty}^{+\infty} g(x)f_X(x)dx; \qquad \text{for a continuous case,} \tag{2.11}$$

and

$$\mathbb{E}[g(X)] = \sum_x g(x)\Pr(X = x) = \sum_x g(x)p(x); \qquad \text{for a discrete case.} \tag{2.12}$$

We can deduce that the mean and variance are special cases of expected values, where the mean is found by taking $g(X) = X$, so that

$$\mu_X = \mathbb{E}[X] = \int_{-\infty}^{+\infty} xf_X(x)dx, \tag{2.13}$$

and the variance is found by taking $g(X) = (X - \mu_X)^2$, which is defined by

$$\begin{aligned} \sigma_X^2 = \mathbb{E}[(X - \mu_X)^2] &= \int_{-\infty}^{+\infty} (x - \mu_X)^2 f_X(x)dx \\ &= \mathbb{E}[X^2] - \mathbb{E}[X]^2 \\ &= \mathbb{E}[X^2] - \mu_X^2. \end{aligned} \tag{2.14}$$

For two random variables X and Y, the covariance is given by

$$\begin{aligned} \text{Cov}[X,Y] &= \mathbb{E}[(X-\mathbb{E}[X])(Y-\mathbb{E}[Y])] \\ &= \mathbb{E}[XY]-\mathbb{E}[X]\mathbb{E}[Y] \\ &= \mathbb{E}[XY]-\mu_X\mu_Y, \end{aligned} \tag{2.15}$$

which is actually a measure of dependence. It is always zero for independent variables and can also be zero for non linear relationships. Finally, if $\mathbf{X}$ and $\mathbf{Y}$ are two random vectors, the *cross-covariance* is used to refer to the covariance between $\mathbf{X}$ and $\mathbf{Y}$, defined by

$$\begin{aligned} \text{Cov}[\mathbf{X},\mathbf{Y}] &= \mathbb{E}[(\mathbf{X}-\mathbb{E}[\mathbf{X}])(\mathbf{Y}-\mathbb{E}[\mathbf{Y}])^T] \\ &= \mathbb{E}[(\mathbf{X}-\boldsymbol{\mu}_X)(\mathbf{Y}-\boldsymbol{\mu}_Y)^T]. \end{aligned} \tag{2.16}$$

2.1.8 Optimisation theory

In most cases numerical techniques are required to search for optima of functions so as to select the best value from some set of possible choices. Here I review some basic principles of optimisation theory along with the Automatic Differentiation Model Builder (ADMB) and simulated annealing methods for solving optimisation problems.

2.1.8.1 Concavity and Convexity

Let us assume that $f(x)$ is a continuous function in the interval $[x_1, x_2]$. This function is *concave* $\Leftrightarrow$

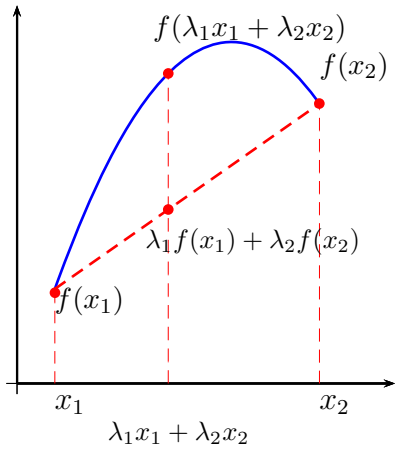
$\forall x_1, x_2 \in \mathbb{R}$ and $\forall \lambda_1, \lambda_2 \in [0,1]$ given that $\lambda_1 + \lambda_2 = 1$, we obtain

$$f(\lambda_1 x_1 + \lambda_2 x_2) \geq \lambda_1 f(x_1) + \lambda_2 f(x_2). \tag{2.17}$$

Figure 2.2 is a graphical representation of a concave function, which shows that for any point falling between x_1 and x_2, the corresponding function value is larger than the value belonging to the chord . Moreover, the function $f(x)$ is strictly concave iff the inequality is strict. Similarly, we define the *convex* function in the form

$$f(\lambda_1 x_1 + \lambda_2 x_2) \leq \lambda_1 f(x_1) + \lambda_2 f(x_2). \tag{2.18}$$

Figure 2.2: Definition of concavity.



2.1.8.2 Jensen's Inequality

For a continuous concave function $f(x) \in \mathbb{R}$ and X a random variable with values $\in \mathbb{R}$, we can write

$$f(\mathbb{E}[X]) \geq \mathbb{E}[f(x)]. \tag{2.19}$$

An important conclusion that is worth mentioning at this juncture is that the local minimum of a convex function is a global minimum too; in addition, the local maximum of a concave function is also a global maximum.

2.1.8.3 Automatic Differentiation Model Builder

Fournier et al. (2012) developed the automatic differentiation model builder (ADMB) programming framework based on automatic differentiation (AD) to numerically compute derivatives of highly nonlinear models with a large number of parameters. It takes advantage of C++ class structures to collect intermediate results and perform internal calculations that implement the AD algorithm, which is based on the chain rule. ADMB uses the reverse AD method as a general strategy for calculating first-order derivatives rather than applying numerical derivative calculation based on finite differences. AD involves evaluating the objective function, storing in memory the value of each intermediate quantity $(t_1, t_2, \ldots, t_n)$ and then apply the chain rule, in which no derivatives are calculated. The parameter estimation is based on maximizing the log-likelihood function using a quasi-Newton algorithm with derivatives obtained using the AD method (ADMB actually minimizes the objective function so in practice the negative log-likelihood is used). ADMB partitions the model specification into three

logical steps: (1) read in the data; (2) declare model parameters; and (3) code the negative log-likelihood function to be minimized with respect to the model parameters.

2.1.8.4 Simulated Annealing

Simulated annealing (SA) is an optimisation technique used to search for the global optimum of a given function. It was inspired from thermodynamics by analogy with annealing of a metal that consists of heating the solid state metal to a high temperature and then cooling it down very slowly (annealing) to ensure thermal equilibrium (Metropolis et al., 1953). This leads to a state with lower energy than the energy of the metal before heating it. In general, SA algorithms are better than greedy algorithms (e.g. Nelder-Mead, EM, Newton-Raphson) because they tend to find the global minimum among many local minima, providing a sufficiently slow cooling technique. The algorithm of the SA can be described as follows:

1. Starting from a very high 'temperature'.
2. Perturb the placement through a defined move.
3. Calculate the change in the score due to the move made.
4. Depending on the change in score, accept or reject the move. The probability of acceptance depends on the current 'temperature'.
5. Update the temperature value by lowering the temperature. Then go back to Second step.

This process is carried out until the 'Freezing Point' is reached. one should probably run this algorithm multiple times, with different initial values, to check whether it finds approximately the same maximum likelihood estimate on each run.

2.1.9 Probability distributions

In this section, I review two distributions of the exponential family, namely the Gaussian and Gamma distributions.

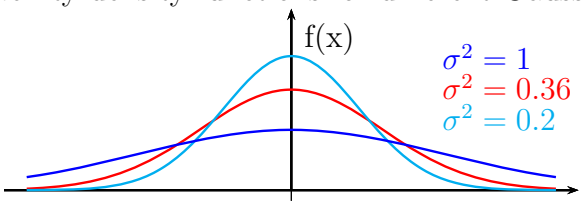
2.1.9.1 The Gaussian distribution

The Gaussian distribution was invented by Carl Friedrich Gauss (1777-1855) and has been used widely to model a distribution of continuous variables. It has been known by the name *Normal*, because most of the physical processes often have distributions that are nearly normal (Lyon, 2014). For the univariate case, the Gaussian distribution can be written in the form

$$\mathcal{N}(x|\mu,\sigma^2) = \frac{1}{\sqrt{2\pi\sigma^2}}\exp\{-\frac{1}{2\sigma^2}(x-\mu)^2\}. \tag{2.20}$$

This distribution is controlled by two parameters which are the mean μ and variance σ^2. One might use the following notation $X \sim \mathcal{N}(\mu,\sigma^2)$ to declare that

Figure 2.3: Probability density functions for different Gaussian distributions.



a random variable X follows a Gaussian distribution. Figure 2.3 illustrates three different Gaussian distributions plotted using different variances.
Assuming that $\mathbf{X}$ is a D-dimensional vector of continuous variables written as $\mathbf{X} = [X_1, X_2, \ldots, X_D]$ respectively. The multivariate Gaussian distribution takes the form

$$\mathcal{N}(\mathbf{X}|\boldsymbol{\mu},\Sigma) = \frac{1}{(2\pi)^{D/2}\sqrt{|\Sigma|}}\exp\left\{-\frac{1}{2}\begin{pmatrix} X_1-\mu_1 \\ X_2-\mu_2 \\ \vdots \\ X_D-\mu_D \end{pmatrix}^T \Sigma^{-1} \begin{pmatrix} X_1-\mu_1 \\ X_2-\mu_2 \\ \vdots \\ X_D-\mu_D \end{pmatrix}\right\}, \tag{2.21}$$

where $\boldsymbol{\mu}$ is a D-dimensional mean vector, and Σ is a $D \times D$ covariance matrix given by

$$\Sigma = \begin{pmatrix} \Sigma_{11} & \cdots & \Sigma_{1D} \\ \Sigma_{21} & \cdots & \Sigma_{2D} \\ \vdots & \ddots & \vdots \\ \Sigma_{D1} & \cdots & \Sigma_{DD} \end{pmatrix} = \begin{pmatrix} \mathrm{Cov}[X_1,X_1] & \cdots & \mathrm{Cov}[X_1,X_D] \\ \mathrm{Cov}[X_1,X_2]^T & \cdots & \mathrm{Cov}[X_2,X_D] \\ \vdots & \ddots & \vdots \\ \mathrm{Cov}[X_1,X_D]^T & \cdots & \mathrm{Cov}[X_D,X_D] \end{pmatrix}. \tag{2.22}$$

2.1.9.2 The Gamma distribution

The gamma distribution was formally introduced by Karl Pearson during the late 19^{th} century to measure the degree of skewness of a continuous random variable. It is well suited for many applications (e.g. financial modelling) and most commonly in Bayesian statistics, where the inverted form of the gamma distribution serves as a conjugate prior for the variance of the Gaussian distribution. Carvalho et al. (2010) and Prado and Lopes (2010) have developed inference techniques to retrieve information from models with fat-tailed noise defined using the inverse gamma distribution. The probability density function of a gamma distribution, is given by

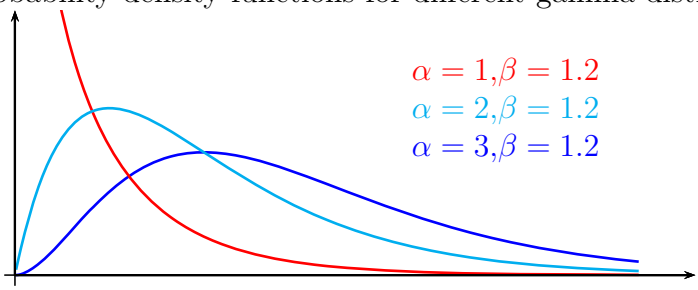
$$f(X|\alpha,\beta) = \frac{\beta^{\alpha}}{\Gamma(\alpha)} X^{\alpha-1} \exp(-\beta X), \tag{2.23}$$

where X is a random variable, α is the shape parameter, and β is the inverse scale parameter, such that $\alpha, \beta, X > 0$. Additionally, the gamma function evaluated at α is defined as

$$\Gamma(\alpha) \equiv \int_0^{\infty} u^{\alpha-1} \exp(-u) du. \tag{2.24}$$

Figure 2.4 illustrates three possible shapes of the probability density function

Figure 2.4: Probability density functions for different gamma distributions.



obtained by varying α and β respectively. The moments, mean and variance, of this distribution are given by $\mathbb{E}[X] = \alpha/\beta$ and $\mathbb{V}\text{ar}[X] = \alpha/\beta^2$. However, the inverse gamma (IG) distribution is defined by the distribution of $Y = 1/X$ where $X \sim \text{G}(\alpha, \beta)$. The resulting density function becomes

$$f(X|\alpha,\beta) = \frac{\beta^{\alpha}}{\Gamma(\alpha)} X^{-(\alpha-1)} \exp(-\beta/X), \tag{2.25}$$

where the moments are given by

$$\mathbb{E}[X] = \beta/(\alpha-1) \quad \text{and} \quad \mathbb{V}\text{ar}[X] = \beta^2/(\alpha-1)^2(\alpha-2).$$

2.1.10 Sufficient Statistic

The concept of sufficient statistic is useful to summarise the data without losing any information; in many practical examples it may be difficult to retain the data due to disk storage space or significant degradation in performance. The formal definition of sufficient statistic can be defined as: let $x_1, \ldots, x_n$ be a random sample from a population depending on θ. A statistic $T = T(x_1, \ldots, x_n)$ is said to be sufficient for the family of probability distributions if the conditional distribution of $x_1, \ldots, x_n$ given $T = t$ is independent of θ. For example, if we observe a random sample $X = \{0, 0, 1, 0, 0, \ldots, 1\}$ generated from a Bernoulli distribution $B(n, p)$ and containing n independent non-identical trials; the sum of the n elements of X entirely depends on the number of 1s that constitutes a sufficient statistic for p.

2.1.11 Prior distribution

The prior distribution is an integral part of Bayesian inference because it serves to define the overall model when it is combined with the likelihood. For instance, the subjectivist approach enables us to develop predictive models as representation of beliefs before we observe the data. The overall model represents an updated form of the degree of belief about observables that we made initially and hence allows the model to learn from experience. What causes Bayesian inference to become widely adopted in statistics and Machine Learning is its capability of transforming the inference problem to be completely dependent of the probability theory. Broadly speaking, the prior is more influential on the posterior distribution when we have few observations to update our beliefs. Below I describe the *subjective*, *objective* and *hierarchical* priors.

2.1.11.1 Subjective prior

The subjective Bayesian approach attempts to extrapolate our prior beliefs to anticipate the future behaviour of physical systems (e.g. weather forecasts). This knowledge can be gained through different means such as results obtained from previous similar experiments, historical data or knowledge obtained from expert systems. Although the subjectivist approach has been widely used for solving many important practical problems, it is still open for many criticisms due to its subjective viewpoint itself. Typically, one may find it very difficult to interpret the

complete detailed specification of beliefs about observables into a language which allows for precise and rigorous analysis. A commonly used method to provide a convenient analytical form rather than a realistic treatment is the *exponential family*, which is not suitable for problems with latent variables.

2.1.11.2 Objective prior

The objective prior distribution approach attempts to apply non-informative priors to problems with limited background knowledge. This prior ignorance is also applicable in instances when one may not be able to represent beliefs about observables into a mathematical language for the prior. Moreover, it attempts to reduce the impact of the prior selection to avoid misleading the overall model, and hence handing over the inference process to the data. This is often said 'letting the data speak for themselves'.
The most commonly used non-informative prior is Jeffrey's prior (Jeffreys, 1946), which takes the form

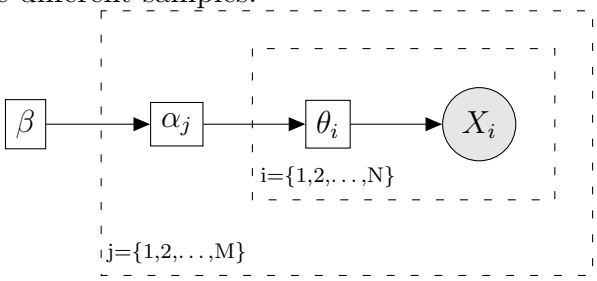
$$\pi_{\boldsymbol{\theta}}(\boldsymbol{\theta}) \propto \det \sqrt{\mathbf{I}(\boldsymbol{\theta})}, \tag{2.26}$$

where $\mathbf{I}(\boldsymbol{\theta})$ is the Fisher information matrix for any given value of $\boldsymbol{\theta}$, given by

$$\mathbf{I}(\boldsymbol{\theta}) = \mathbb{E}_{\boldsymbol{\theta}} \left[\frac{\partial^2}{(\partial \boldsymbol{\theta})^2} \log(p(x|\boldsymbol{\theta})) \right]. \tag{2.27}$$

The expectation is determined with respect to the probability function $p(x|\boldsymbol{\theta})$. The Jeffrey's prior is valid as long as $\mathbf{I}(\boldsymbol{\theta})$ is defined and positive definite. For example, in case that the data follows a Gaussian distribution $\mathcal{N}(\mu, \sigma^2)$, the general rule identifies a uniform prior on μ while fixing σ, whereas it identifies a prior $\pi(\sigma) \propto 1/\sigma$ while fixing μ. However, the Jeffrey's multivariate rule yields to $\pi(\mu, \sigma) \propto 1/\sigma^2$. This is also known as an improper prior because the area of a uniform distribution varying between $(-\infty, +\infty)$ will achieve a sum greater than one. Berger et al. (2009) proposed an enhanced version of the Jeffrey's prior that accounts for high dimensional problems, known as the *Reference priors*.

Figure 2.5: A simple generative model for Bayesian inference with hierarchical prior structure. The dashed rectangles denote a plate representation, that is, a repetition over the different samples.



2.1.11.3 Hierarchical prior

The hierarchical approach consists of splitting the prior into multiple levels of hyper-prior distributions, which are usually flat and non-informative. An unnormalised posterior distribution can be written in the form

$$p(\boldsymbol{\theta}|\mathbf{X}) \propto p(\mathbf{X}|\boldsymbol{\theta})p(\boldsymbol{\theta}).$$

However, if we believe that the parameter $\boldsymbol{\theta}$ depends on a hyper-prior distribution of $\boldsymbol{\alpha}$, a hierarchical structure replaces the prior $p(\boldsymbol{\theta})$ into a likelihood $p(\boldsymbol{\theta}|\boldsymbol{\alpha})$ and a prior $p(\boldsymbol{\alpha})$, so that

$$p(\boldsymbol{\theta}, \boldsymbol{\alpha}|\mathbf{X}) \propto p(\mathbf{X}|\boldsymbol{\theta})p(\boldsymbol{\theta}|\boldsymbol{\alpha})p(\boldsymbol{\alpha}).$$

Additionally, if we believe that the parameter $\boldsymbol{\alpha}$ depends on another hyperprior distribution $\boldsymbol{\beta}$, the posterior would then become

$$p(\boldsymbol{\theta}, \boldsymbol{\alpha}, \boldsymbol{\beta}|\mathbf{X}) \propto p(\mathbf{X}|\boldsymbol{\theta})p(\boldsymbol{\theta}|\boldsymbol{\alpha})p(\boldsymbol{\alpha}|\boldsymbol{\beta})p(\boldsymbol{\beta}).$$

This hierarchical structure could continue further until estimating the optimal prior distributions from the data. Hence the hierarchical prior structure can be represented in the form

$$\begin{aligned} p(\boldsymbol{\theta}) &= \int p(\boldsymbol{\theta}|\boldsymbol{\alpha})p(\boldsymbol{\alpha})d\boldsymbol{\alpha}, \\ &= \int p(\boldsymbol{\theta}|\boldsymbol{\alpha})d\boldsymbol{\alpha} \int p(\boldsymbol{\alpha}|\boldsymbol{\beta})p(\boldsymbol{\beta})d\boldsymbol{\beta}. \end{aligned} \tag{2.28}$$

Figure 2.5 illustrates a three stage hierarchical model such that at the first stage the data X_i is assumed to be sampled from a certain distribution with a parameter θ_i, at the second stage the between data variation is modelled with α_j, and at the third stage, one can set a hyperprior distribution β on the α_js.

2.1.12 Conjugate-exponential models

Conjugate-exponential models involve distributions that belong to the same exponential family where the prior is chosen to have similar structure to the likelihood. Accordingly, both prior and likelihood would become conjugate, and the prior would be called a conjugate prior with respect to the likelihood (Bernardo and Smith, 2000). The exponential family of distributions defined over a random sample, $X = \{x_1, x_2, \ldots, x_n\}$, and a set of parameters $\boldsymbol{\theta}$ can be described in the form

$$p(X|\boldsymbol{\theta}) = \prod_{j=1}^{n} f(x_j)[g(\boldsymbol{\theta})]^n \exp\left\{\sum_{i=1}^{k} c_i\phi_i(\boldsymbol{\theta})\left(\sum_{j=1}^{n} h_i(x_j)\right)\right\}, \tag{2.29}$$

where $\boldsymbol{\theta}$ is a vector of *natural parameters* of the distribution; $\phi_i(\boldsymbol{\theta})$ is a function of $\boldsymbol{\theta}$; the elements of $\mathbf{h}(x) = \left[\sum_{j=1}^{n} h_1(x_j), \ldots, \sum_{j=1}^{n} h_k(x_j)\right]$ are functions of x that provides the *sufficient statistics* of the distribution; c_i is a constant term; $f(x_j)$ represents a function over the variable X, and $g(\boldsymbol{\theta})$ is the normalisation term. A conjugate prior for the likelihood function described in Equation (2.29) can be written in the form

$$p(\boldsymbol{\theta}|\boldsymbol{\tau}) = \frac{g(\boldsymbol{\theta})^{\tau_0} \exp\left\{\sum_{i=1}^{k} c_i\phi_i(\boldsymbol{\theta})\tau_i\right\}}{\mathcal{Z}(\boldsymbol{\tau})}, \qquad \text{given that} \quad \boldsymbol{\theta} \in \Theta, \tag{2.30}$$

where $\boldsymbol{\tau} = \{\tau_0, \tau_1, \ldots, \tau_k\}$ is the set of hyper-parameters, and the normalisation constant represented in the denominator is given by

$$\mathcal{Z}(\boldsymbol{\tau}) = \int g(\boldsymbol{\theta})^{\tau_0} \exp\left\{\sum_{i=1}^{k} c_i\phi_i(\boldsymbol{\theta})\tau_i\right\} d\boldsymbol{\theta}.$$

Hence, the posterior will have the same functional form as the prior, given by

$$p(\boldsymbol{\theta}|\mathbf{X}) = \frac{\prod_{j=1}^{n} f(x_j)[g(\boldsymbol{\theta})]^{n+\tau_0} \exp\left\{\sum_{i=1}^{k} c_i\phi_i(\boldsymbol{\theta})\left(\sum_{j=1}^{n} h_i(x_j) + \tau_i\right)\right\}}{\int g(\boldsymbol{\theta})^{\tau_0} \exp\left\{\sum_{i=1}^{k} c_i\phi_i(\boldsymbol{\theta})\tau_i\right\} d\boldsymbol{\theta}}. \quad (2.31)$$

Many common distributions such as the Gaussian, Dirichlet, Beta, Gamma, Multinomial and Bernoulli are members of this family. For example, the conjugate prior for the normal likelihood is the normal gamma prior. However, from a practical point of view, the more complex the prior function we choose, the more computationally intensive the exploration of the posterior would become.

2.2 Graphical Models

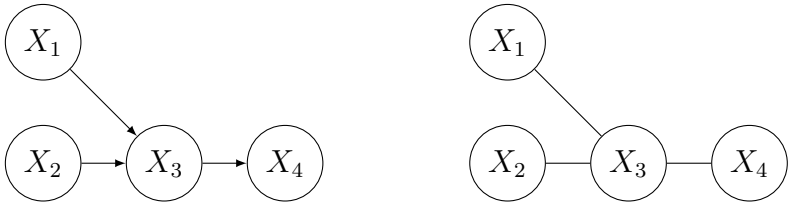
In this section I introduce some basic concepts of graphical models, which can be used for building probabilistic models and inference algorithms. Moreover, I describe the two main members of the graphical models, *directed* and *undirected* graphs, and some of their properties.

2.2.1 Basic concepts

A graph G is defined by a set of nodes (or vertices) and links (or edges). The former can be used to describe a set of random variables $\mathbf{X} = (X_1, X_2, \ldots, X_n)$; however, the latter consists of a set of links used to describe a probabilistic relationship among these variables $L = (l_1, l_2, \ldots, l_m)$. A graphical model G can be described using the following notation: $G = (\mathbf{X}, L)$.
The name of directed and undirected graphs are derived from the types of links used in a graph. A graph is called directed when all the links used in a graph are directed; whereas it is called undirected when all links are undirected. Examples of directed and undirected graphs are illustrated in Figures 2.6 (a) and (b), respectively.

2.2.2 Directed Graphs

An important property of directed graphs is the concept of parents and children; for example, if there is a directed link from X_i to X_j (i.e. $X_i \rightarrow X_j$) we can say that X_i is the parent of X_j, and X_j is the child of X_i. By referring to the



(a) Directed Graph (b) Undirected Graph

Figure 2.6: Examples of directed and undirected graphs.

Figure 2.6 (a) we can say that the nodes X_1 and X_2 are the parents of X_3, and X_4 is the only child of X_3. A graph G that is composed of directed edges between nodes such that there is no way to find a closed path or cycle while following a consistently-directed sequence of edges is called a *direct acyclic graph* (DAG); these graphs are also known as belief networks or Bayesian networks. The joint probability distribution for a DAG defined over a set of n variables can be represented as

$$p(X_1, X_2, \ldots, X_n) = \prod_{i=1}^{n} p(X_i|\text{pa}(X_i)), \tag{2.32}$$

where $\text{pa}(X_i)$ are the parents of X_i. We can then write the joint distribution of the graphical model (Figure 2.6 (a)) as a product of conditional distributions, such that

$$p(X_1, X_2, X_3, X_4) = p(X_1)p(X_2)p(X_3|X_1, X_2)p(X_4|X_3). \tag{2.33}$$

The joint distribution defines the probability of events in terms of random variables, which is often used to refer to a probability density function when observing continuous variables. By marginalising both sides over X_3 and X_4, we obtain

$$p(X_1, X_2) = p(X_1)p(X_2), \tag{2.34}$$

This equation tells us that the knowledge about X_2 does not improve our knowledge about X_1, and vice-versa. It turns out to represent a joint distribution of independent random variables by the product of their marginals. The variables X_1 and X_2 are said to be independent, if

$$p(X_2|X_1) = p(X_2). \tag{2.35}$$

By referring back to the conditional distribution described in section 2.1.3, we can write

$$p(X_2|X_1) = \frac{p(X_1, X_2)}{p(X_1)}. \tag{2.36}$$

One way to solve this equation is to reverse the problem by propagating the observed evidence back to the opposite direction. The Equation (2.36) can then be expressed as follows

$$p(X_2|X_1) = \frac{p(X_1|X_2) \times p(X_2)}{p(X_1)}, \tag{2.37}$$

which reminds us about the Bayes theorem, described previously in section 2.1.4. The prior $p(X_2)$ describes our initial belief about X_2 before observing the evidence, the likelihood $p(X_1|X_2)$ is determined by our model, the evidence $p(X_1)$ is the normalisation constant, and finally the posterior $p(X_2|X_1)$ is our new knowledge about X_2 obtained when observing X_1. Therefore graphical models can be adopted as a convenient method for graphically representing a family of probability distributions over a large number of random variables.

2.2.3 Undirected Graphs

In this section, I briefly introduce undirected graphs, which is the second largest family of graphical models. They are called undirected graphs because the links do not carry arrows and do not provide any direction, as illustrated in Figure 2.6 (b). Unlike belief networks, this type of graphs use an alternative factorisation method to describe the joint probability distribution, which takes the form

$$p(X_1, X_2, \ldots, X_n) = \frac{1}{\mathcal{Z}} \prod_{c}^{C} \phi_c(x_c). \tag{2.38}$$

where $\mathcal{Z}$ is the normalisation constant, and $\phi_c(X_c)$ is the potential functions over the maximal *cliques* of the graph. A clique is a subset of fully connected nodes defined for the elements within the clique. For instance, if we factorise the joint probability distribution of the graph that is illustrated in Figure 2.6 (b), we obtain

$$p(X_1, X_2, X_3, X_4) = \frac{1}{\mathcal{Z}} \phi_c(X_1, X_3) \phi_c(X_2, X_3) \phi_c(X_3, X_4), \tag{2.39}$$

where $\phi_c(X_1, X_3)$, $\phi_c(X_2, X_3)$ and $\phi_c(X_3, X_4)$ are the potential functions over the maximal cliques of the graph. Moreover, this type of graphs is also known as a *Markov network* because the set of random variables have a Markov property.

2.3 Scoring functions

In this section I review the AIC, AICc and DIC methods for comparing the fits of models on a given outcome. These methods are in favour of selecting the model with the mimimum score so as to approximate the underlying process that has generated the observed data.

2.3.1 Akaike Information Criterion

The Akaike information criterion (AIC) statistic (Akaike, 1973) is a method used to select a model from a set of models; it penalises the likelihood for the number of parameters estimated, such that

$$\text{AIC} = -2\mathcal{L}(\hat{\theta}) + 2D, \tag{2.40}$$

where $\mathcal{L}(\hat{\theta})$ is the log-likelihood of the model evaluated at the maximum likelihood estimate of θ and D is the number of (independent) model parameters. This AIC statistic indicates that the smaller the value, the better the model.

2.3.2 Akaike Information Criterion corrected

An alternative version of the AIC statistic with a more severe penalty for the number of parameters estimated is known as the bias-corrected AIC, denoted by AICc (Hurvich and Tsai, 1989), defined as

$$\text{AICc} = -2\mathcal{L}(\hat{\theta}) + 2D\frac{n}{n - D - 1}. \tag{2.41}$$

This criterion avoids overfitting by replacing the penalty term of AIC with an exact expression for the bias adjustment and provides improved model selection for small samples. However, as n gets large, AICc converges to AIC: rendering the AICc a more effective statistic in practice.

2.3.3 Deviance Information Criterion

DIC (Spiegelhalter et al., 2002) is a Bayesian version of AIC (Akaike, 1973) intended for hierarchical Bayesian models that consists of replacing the maximum likelihood estimate with the posterior mean and replacing the number of parameters estimated in the model with a data-based bias correction, such that

$$DIC = \bar{D} + p_D. \tag{2.42}$$

The first term defines the posterior expectation of the deviance

$$\bar{D} = E_{\theta|y}[D(\theta)] = E_{\theta|y}[-2\ln f(y|\theta)]. \tag{2.43}$$

This is a function of -2 times log-likelihood and it attains smaller values for better fitting models.
The second term is the effective number of parameters that measures the complexity of the model, defined as the difference between the posterior mean of the deviance and the deviance evaluated at the posterior mean $\bar{\theta}$ of parameters, such that:

$$p_D = \bar{D} - D(\bar{\theta}) = E_{\theta|y}[D(\theta)] - D(E_{\theta|y}[\theta]) = E_{\theta|y}[-2\ln f(y|\theta)] + 2\ln f(y|\bar{\theta}). \tag{2.44}$$

p_D measures the effective number of parameters.

2.4 Statistical Inference

In this section I explore relevant statistical methods that are used in the course of my research so as to provide a basic understanding of the specific problem. There are two general approaches to solving statistical problems: the first is based on the frequentist paradigm, and the second is based on Bayesian methods. Problems involving reasonably large datasets are well suited to frequentist methods; however, Bayesian methods are desirable for problems involving small datasets, missing values and where past knowledge of similar experiments are available (Wakefield, 2013, page 144).

2.4.1 Frequentist Statistical Methods

Under the frequentist approach to inference, parameters and hypotheses are viewed as unknown but fixed quantities; whereas probabilities are related to frequencies of events. Ronald Fisher invented the so-called maximum likelihood estimation (MLE) method in the mid 1920s so as to determine the parameter values that maximise the likelihood function, given a model $\mathcal{M}$. This estimator is found to be important because it satisfies the following properties: consistency, normality, and efficiency.

Consider a set of independent identically distributed (i.i.d.) samples denoted by $X = \{x_1, \ldots, x_n\}$ for which we aim to estimate the probability distribution over the data, $p(X|\boldsymbol{\theta})$; one approach consists of involving the MLE to fit model parameters such that

$$\boldsymbol{\theta}_{ML} = \arg\max_{\boldsymbol{\theta}} p(x|\boldsymbol{\theta}, \mathcal{M}), \tag{2.45}$$

The likelihood function, which is the probability of data given parameters, for i.i.d. samples can be written as the product of individual observations in the data set such that

$$p(X|\boldsymbol{\theta}) = \prod_{i=1}^{n} p(x_i|\boldsymbol{\theta}), \tag{2.46}$$

Now let us consider an example where the data is derived from an unknown Gaussian distribution $\mathcal{N}(\mu, \sigma^2)$ where we would like to infer the model parameters $\boldsymbol{\theta} = \{\mu, \sigma^2\}$ through maximum likelihood. By assuming that the i.i.d. condition is still valid, the log-likelihood function can be written as

$$\mathcal{L}(\mu, \sigma^2) \equiv \sum_{i=1}^{n} \log p(x_i|\mu, \sigma^2) = -\frac{n}{2}\log(2\pi) - \frac{n}{2}\log(\sigma^2) - \frac{1}{2}\sum_{i=1}^{n}(x_i - \mu)^2. \tag{2.47}$$

By taking the partial derivative for each parameter and equating it to zero, we obtain the MLE for model parameters, such that

$$\mu_{\text{ML}} = \frac{1}{n}\sum_{i=1}^{n} x_i \qquad \text{and} \qquad \sigma^2_{\text{ML}} = \frac{1}{n}\sum_{i=1}^{n}(x_i - \mu_{\text{ML}})^2. \tag{2.48}$$

So far we demonstrated the maximum likelihood as a simple method for estimating the parameter values of a model. But does it work well in practice?

Let us assume that we have been given a task to judge whether a particular coin is biased or not. Accordingly, we toss the coin few times (e.g. four times), and

in each attempt it lands tails. What conclusion can we deduce?
Based on these events, the probability of observing a tail is always one, and hence we would be very confident, based on maximum likelihood estimation, to announce that the coin is biased because there is no evidence at all for observing heads. But does this convince the audience?
One would not believe the coin is biased just by observing a sequence of four tails. However, it would be more conceivable to convince the audience if we observed a sequence of hundred tails perhaps, or couple of heads in a very long sequence. We can deduce that the maximum likelihood is not suitable for problems with small dataset size; but it can be a good choice for analysing models with few parameters and a big dataset. Accordingly, we conclude that MLE is not a satisfactory general solution to infer model parameters because of over fitting problems when applied to models with many parameters and few observations. One might propose incorporating prior belief about the model, $p(\boldsymbol{\theta})$, making it possible to determine the *maximum a posteriori* (MAP) parameters by using Bayes' theorem, which can be written in the form

$$\boldsymbol{\theta}_{\mathrm{MAP}} = \arg\max_{\boldsymbol{\theta}}(p(x|\boldsymbol{\theta}, \mathcal{M})p(\boldsymbol{\theta}|\mathcal{M})). \tag{2.49}$$

This can be interpreted as a summary of the posterior. $\boldsymbol{\theta}_{\mathrm{MAP}}$ provides an exact point estimate for the parameter of interest instead of providing a whole distribution of possible values.

2.4.2 Bayesian Inference

The problem of learning is often decomposed into three levels of inference (Rattray, 2008): at the first level of inference, we assume that uncertainties can be represented by probabilistic models; at the second level of inference, we assume that our proposed mathematical model is correct and then we fit it to the data to estimate the model parameters; at the third level of inference, we compare different plausible models and select the one that best represents the data, known as model selection. The selected model will be our best choice for representing uncertainties in decisions. All these levels of inference can be solved by using Bayes theorem, which is described in section 2.1.4. The advantage of using Bayesian inference over the frequentist method is its ability to allow inference in difficult conditions, especially when the data set is limited, noisy and contains missing

information; but its difficulty lies in specifying the prior distribution because an incorrect prior distribution in a small data set will have a great distorting effect on the results. The model parameters $\boldsymbol{\theta}_i$ in a Bayesian model $\mathcal{M}_i$ are estimated by invoking Bayes' rule, such that

$$\underbrace{p(\boldsymbol{\theta}_i|\mathbf{X}, \mathcal{M}_i)}_{Posterior} = \frac{\overbrace{p(\mathbf{X}|\boldsymbol{\theta}_i, \mathcal{M}_i)}^{Likelihood}\overbrace{p(\boldsymbol{\theta}_i|\mathcal{M}_i)}^{Prior}}{\underbrace{p(\mathbf{X}|\mathcal{M}_i)}_{Evidence}}. \tag{2.50}$$

where $\mathcal{M}_\mathbf{i}$ is a particular model from a set of models $\{\mathcal{M}_1, \mathcal{M}_2, \ldots, \mathcal{M}_N\}$, and $\boldsymbol{\theta}_i$ is the corresponding parameter vector. In Bayesian modelling we first need to define the likelihood function $p(\mathbf{X}|\boldsymbol{\theta}_i, \mathcal{M}_i)$, then express our prior beliefs about the parameter values before seeing the data, whereas the evidence $p(\mathbf{X}|\mathcal{M}_i)$ (or the marginal likelihood) involves an integration over all possible parameters of the model, such that

$$p(\mathbf{X}|\mathcal{M}_i) = \int_{\boldsymbol{\theta}_i \in \Theta} p(\mathbf{X}|\boldsymbol{\theta}_i, \mathcal{M}_i) p(\boldsymbol{\theta}_i|\mathcal{M}_i) d\boldsymbol{\theta}_i, \tag{2.51}$$

where $\mathbf{X}$ is a set of random variables. After observing some data, one would be able to compute the posterior distribution $p(\boldsymbol{\theta}_i|\mathbf{X}, \mathcal{M}_i)$ that can be used to improve our knowledge about the parameters. Also, the Bayesian inference allows the estimation of latent or hidden variables of a model given observed data and a particular parameter setting, so that

$$p(\mathbf{Z}|\mathbf{X}, \boldsymbol{\theta}_i, \mathcal{M}_i) = \frac{p(\mathbf{X}|\mathbf{Z}, \boldsymbol{\theta}_i, \mathcal{M}_i) p(\mathbf{Z}|\boldsymbol{\theta}_i, \mathcal{M}_i)}{p(\mathbf{X}|\boldsymbol{\theta}_i, \mathcal{M}_i)}. \tag{2.52}$$

where $\mathbf{X}$ is a set of real-valued observations with D dimensions, and $\mathbf{Z}$ is a vector of latent variables which lies in a lower dimensional space. The probability of the observed data for a particular parameter setting can be obtained by integrating out over all latent variables, given by

$$p(\mathbf{X}|\boldsymbol{\theta}_i, \mathcal{M}_i) = \int p(\mathbf{Z}|\boldsymbol{\theta}_i, \mathcal{M}_i) p(\mathbf{X}|\mathbf{Z}, \boldsymbol{\theta}_i, \mathcal{M}_i) d\mathbf{Z}. \tag{2.53}$$

In order to obtain the complete marginal likelihood one should integrate over all possible sets of parameters and hidden variables rather than optimising them,

which can be described in the form

$$p(\mathbf{X}|\mathcal{M}_i) = \int_{\boldsymbol{\theta}_i \in \Theta} p(\boldsymbol{\theta}_i|\mathcal{M}_i) \int p(\mathbf{Z}|\boldsymbol{\theta}_i, \mathcal{M}_i) p(\mathbf{X}|\mathbf{Z}, \boldsymbol{\theta}_i, \mathcal{M}_i) d\mathbf{Z} d\boldsymbol{\theta}_i. \tag{2.54}$$

In the subsequent sections 2.4.3 and 2.1.11 we shall explain how one could use Bayesian inference to complete model selection and how the choice of prior is important in determining the posterior distribution for predicting future events.

2.4.3 Occam's Razor and Bayesian Model selection

The principle of simplicity dates back to the days of Aristotle, who wrote

> *Nature operates in the shortest way possible.*

Several scientists, philosophers and priests who proceeded Aristotle, provided different point of views. Until the 14^{th} Century when William of Ockham an English Franciscan friar Father stated

> *Pluralitas non est ponenda sine necessitate.*

This statement could be translated as 'Plurality should not be posited without necessity'. Ockham was an important figure in the medieval era, and because he used to cut out or shave away the arguments of others, his principle became known as Ockham's razor or Occam's razor. This principle is still valid till our modern days and has been used by many scientists. Hawking (1995), one of the most brilliant theoretical physicists in our days, stated

> *We could still imagine that there is a set of laws that determines events completely for some supernatural being, who could observe the present state of the universe without disturbing it. However, such models of the universe are not of much interest to us mortals. It seems better to employ the principle known as Occam's razor and cut out all the features of the theory that cannot be observed.*

We stated in section 2.4.1 that a major limitation of the maximum likelihood approach in determining model parameters is due to the problem of over fitting. Bayesian inference is an alternative approach that avoids this problem and consists of computing the posterior distribution over model parameters, which takes the form

$$p(\boldsymbol{\theta}_i|\mathbf{X}, \mathcal{M}_i) = \frac{p(\mathbf{X}|\boldsymbol{\theta}_i, \mathcal{M}_i) p(\boldsymbol{\theta}_i|\mathcal{M}_i)}{p(\mathbf{X}|\mathcal{M}_i)}. \tag{2.55}$$

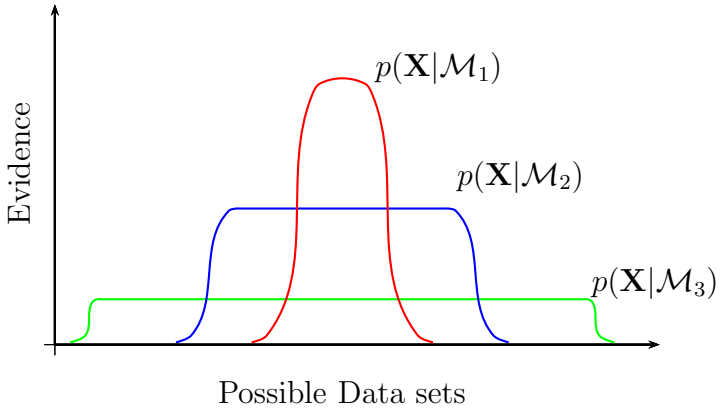
Once determined, this distribution enables us to rectify or correct our prior beliefs over the parameter values after observing the data. The model evidence or the marginal likelihood of the data is described in the denominator of Equation (2.55) that consists of integrating out over all possible parameters settings, as defined in Equation (2.51). We represent it here again for the sake of convenience, so that

$$p(\mathbf{X}|\mathcal{M}_i) = \int_{\boldsymbol{\theta}_i \in \Theta} p(\mathbf{X}|\boldsymbol{\theta}_i, \mathcal{M}_i) p(\boldsymbol{\theta}_i|\mathcal{M}_i) d\boldsymbol{\theta}_i.$$

Finding the marginal likelihood is an important task because on one hand it enables us to compute the posterior distribution and on the other hand it is necessary to develop Bayesian model comparison for finding the model that best describes the data. In this situation we are not going to fit the parameters to the data, but we are going to integrate out over model parameters to avoid the over fitting. This approach does not prevent us from choosing models with infinitely large number of parameters because the size of the complexity penalty increases as we increase the model complexity, as we shall see shortly. Hence the Occam's razor becomes crucial for applying a trade-off for finding the best model.

Figure 2.7 illustrates the Occam's razor axiom where the horizontal axis represents the space of possible data sets to be modelled so that each point on this axis represents a particular data set; the vertical axis represents the normalised distribution of the marginal likelihood, which is integrable to one. A common approach for simulating a data set consists of averaging the probability of the data with respect to the values of the parameters, which are taken from their prior distributions $p(\boldsymbol{\theta}_i|\mathcal{M}_i)$. Accordingly, if a model has low variability, the generated data sets would appear almost with the same pattern —simple representation. On the other hand, if a model has high variability, the generated data sets would then appear to be very different —complex representation. For example, Figure 2.7 illustrates three models $\mathcal{M}_1, \mathcal{M}_2$ and $\mathcal{M}_3$ with increased complexity, such that the first model $p(\mathbf{X}|\mathcal{M}_1)$ represents a very simple representation (because it generates a limited variability of data sets), the third model $p(\mathbf{X}|\mathcal{M}_3)$ represents a very complex representation (because it generates a wide range of data sets); however, the second model $p(\mathbf{X}|\mathcal{M}_2)$ represents a reasonable level of complexity. In general, one may select the model that would provide the highest marginal likelihood value (known as model selection) or estimate some quantity under each candidate model and then construct a weighted average over all of them

Figure 2.7: Pictorial representation of Occam's razor, adapted from (MacKay, 2003).



(known as model averaging). When the computation of the marginal likelihood becomes intractable, one may approximate the problem by choosing *Maximum a Posteriori* (MAP) estimate, as defined in Equation (2.49), given by

$$\boldsymbol{\theta}_{\text{MAP}} = \arg\max_{\boldsymbol{\theta}}(p(x|\boldsymbol{\theta}, \mathcal{M})p(\boldsymbol{\theta}, \mathcal{M})),$$

which is equivalent to work on a simplified form of the posterior, such that

$$\textit{Posterior} \propto \textit{Likelihood} \times \textit{Prior}.$$

This assumes that the posterior distribution is maximised at the point $\boldsymbol{\theta}_{\text{MAP}}$, which is known as MAP solution of the model. The Bayesian Information Criterion (BIC) (Schwarz, 1978) can be obtained from the Laplace approximation applied to the evidence, and so taking logs we obtain

$$\log p(\mathbf{X}) \approx \log p(\mathbf{X}|\boldsymbol{\theta}_{\text{MAP}}) + \log p(\boldsymbol{\theta}_{\text{MAP}}) + \frac{D}{2}\log(2\pi) - \frac{1}{2}\log\det(\mathbf{A}), \quad (2.56)$$

where D is the space dimension of the data set $\mathbf{X}$, and $\mathbf{A}$ is the second order derivative of the posterior, which will be developed later. By assuming that the $\det(\mathbf{A}) \propto n^N$, where n is the size of the data set and N is the number of model parameters, we obtain the BIC expression, which can be written in the form

$$\log p(\mathbf{X}) \approx -2\log p(\mathbf{X}|\boldsymbol{\theta}_{\text{ML}}) + N\log n, \quad (2.57)$$

A similar criterion was developed and known as the Akaike Information Criterion (AIC)

$$\log p(\mathbf{X}) \approx -2 \log p(\mathbf{X}|\boldsymbol{\theta}_{\mathrm{ML}}) + 2n. \tag{2.58}$$

Both criteria penalize the model when the number of parameters increase unnecessarily. A limitation of a AIC and BIC scores is that they do not account for parameter correlations, and hence cannot be used with regularised models (Rattray, 2008).

2.4.4 Bayesian Hierarchical Modelling

Bayesian hierarchical modelling is a statistical model written modularly (or in terms of sub-models) so as to integrate models for both within-unit analysis and across-unit analysis. The aim is to estimate the parameters of the posterior distribution using the Bayesian method. The within-unit model is used to describe the model characteristics over a single unit or population; however, the across-unit analysis is used to account for heterogeneity across all populations (Allenby et al., 2005). Exchangeability is very useful in Bayesian statistics (de Finetti, 1931). A set of random variables $X_1, X_2, \ldots, X_n$ is exchangeable if the joint probability $p(X_1, X_2, \ldots, X_n)$ is invariant to permutation of the indices; that is, for any permutation π,

$$p(X_1, X_2, \ldots, X_n) = p(X_{\pi_1}, X_{\pi_2}, \ldots, X_{\pi_n}). \tag{2.59}$$

Lunn et al. (2013) stated that the exchangeability assumption is equivalent to assuming the observations were independent and identically distributed from a distribution with unknown parameters. Bernardo et al. (1983) showed that one must assume symmetry among the parameters of the prior distribution in case no ordering or grouping of the parameters can be made. This symmetry is represented probabilistically by exchangeability for which there exists a unique probability measure P on [0,1] such that

$$p(X_1, X_2, \ldots, X_n) = \int \Pi_{i=1}^{n} p(X_i|\boldsymbol{\theta}) dP. \tag{2.60}$$

Let x_{ij} be an observation i within a population j, and $\boldsymbol{\theta}_j$ a set of parameters governing the data generating X_j. Assume that the parameters $\boldsymbol{\theta}_1, \boldsymbol{\theta}_2, \ldots, \boldsymbol{\theta}_n$ are generated exchangeably from n populations with distribution governed by a

hyperparameter $\boldsymbol{\phi}$. The structure of the Bayesian hierarchical model described here involves three stages for inference, such that

Stage 1 : $X_j \sim_{ind} p(X_j|\boldsymbol{\theta}_j)$

Stage 2 : $\boldsymbol{\theta}_j|\boldsymbol{\phi} \sim_{i.i.d} p(\boldsymbol{\theta}_j|\boldsymbol{\phi})$

Stage 3 : $\boldsymbol{\phi} \sim \pi(\boldsymbol{\phi})$

In the first stage the data is derived from a distribution that depends on a parameter $\boldsymbol{\theta}$ such that each population X_j is controlled by its own parameter $\boldsymbol{\theta}_j$. At the second stage of the hierarchy, the parameter $\boldsymbol{\theta}_j$ comes from a common distribution that depends on a hyperparameter $\boldsymbol{\phi}$. Finally, at the third stage of the hierarchy we place a prior on the hyperparameter $\boldsymbol{\phi}$. Only $\boldsymbol{\phi}$ has a prior that is set manually. Thus, the joint posterior distribution of interest in hierarchical models can be written as

$$p(\boldsymbol{\theta}, \boldsymbol{\phi}|\mathbf{X}) \propto p(\mathbf{X}|\boldsymbol{\theta}, \boldsymbol{\phi})p(\boldsymbol{\theta}, \boldsymbol{\phi}) = p(\mathbf{X}|\boldsymbol{\theta})p(\boldsymbol{\theta}|\boldsymbol{\phi})\pi(\boldsymbol{\phi}). \tag{2.61}$$

Alternatively, the marginal posteriors can be written as

$$p(\boldsymbol{\theta}|\mathbf{X}) = \int p(\boldsymbol{\theta}, \boldsymbol{\phi}|\mathbf{X})d\boldsymbol{\phi} \qquad \text{or} \qquad p(\boldsymbol{\phi}|\mathbf{X}) = \int p(\boldsymbol{\theta}, \boldsymbol{\phi}|\mathbf{X})d\boldsymbol{\theta}. \tag{2.62}$$

2.5 Bayesian Networks

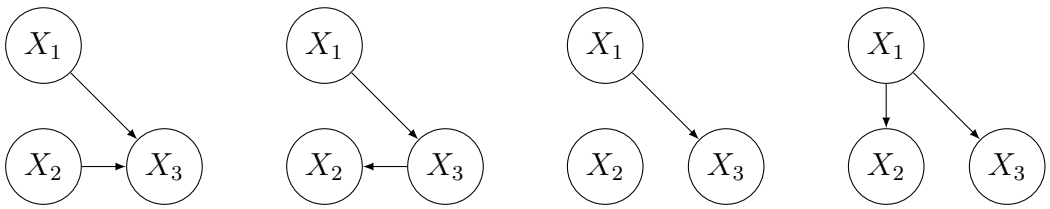
A Bayesian network (BN) is a special type of probabilistic graphical model that explicitly represents conditional independence relationships among random variables via a directed acyclic graph (DAG). For example, Figure 2.6 (a) represents the probabilistic relationships between random variables (X_1, X_2, X_3, X_4) and whose edges correspond to direct influence of one node on another (Koller and Friedman, 2009).

In order to fully specify the Bayesian network, it is necessary to specify the conditional probability distribution for each node upon its parents. Often these conditional distributions include unknown parameters which must be estimated from data in one of the following ways: maximum likelihood approach, expectation-maximisation algorithm or Markov chain Monte Carlo. The latter handles the problem in a Bayesian approach such that all parameters are treated as additional unobserved variables and estimated after approximating the full posterior

distribution over all nodes, which is conditional on the observed data. However, the network structure can be either constructed by expert's knowledge or learned from data. In this thesis I will review basic principles of the *search and score* approach for learning the structure of BNs.

2.5.1 Structure learning

The search and score approach is an optimisation problem that consists of finding a BN that maximises a given scoring function. The vast majority of the search methods used in structure learning are local search procedures such as greedy hill climbing, as described in Algorithm 1. The method consists of searching for a structure that maximises a given score. In case of using the likelihood score, our objective would be to find both a graph $\mathcal{G}$ and the MLE of parameters that maximise the likelihood. Three operations can be used for exploring the search space, which are: (a) reversing an edge, (b) deleting an edge or (c) adding an edge. Any move in the search space is subject to the condition that the resulting structure is a valid BN (DAG). The search evolves in the directions that most increases the score and stops when it does not find a local move that can increase the score. Suppose the original network is the one shown in Figure 2.8(a), the greedy-search procedure starts randomly by reversing the edge between X_2 and X_3, deleting the edge between X_2 and X_3, and finally adding an edge between X_1 and X_2 as it improves the score at each step until it does not find any local move that increases the score.



(a) Original network (b) Reverse (c) Delete (d) Add

Figure 2.8: Examples of a search problem applied on a given network (a) with typical operations: (b) reverse an edge, (c) delete an edge and (d) add an edge.

Algorithm 1 Greedy local search algorithm with search operators, adapted from (Koller and Friedman, 2009)

```
1: procedure GREEDY-LOCAL-SEARCH( σ_0,        //initial candidate solution.
   score,                                     //Score.
   O                          //a set of search operators.  )
2:     σ_best ← σ_0
3:     Progress ← true
4:     while Progress is true do
5:         σ ← σ_best
6:         for each operator o ∈ O do
7:             σ_o ← o(σ) //result of applying o on σ
8:             if σ_o is legal solution  then
9:                 if score(σ_o) > score(σ_best )  then
10:                    σ_best ← σ_o
11:                else
12:                    Progress ← false
13:                end if
14:            end if
15:        end for
16:    end while
17:    return σ_best
18: end procedure
```

2.6 Dynamic Bayesian Networks

A Dynamic Bayesian Network (DBN) is an extension of Bayesian networks that models time series by relating several successive instances (of BNs) through arcs so as to represent how the state of a random variable changes over time Murphy (2002). For a particular case when intra-slice dependencies (connections within time slices) do not exist, one can turn the DBN into a first-order multivariate autoregressive model, which means that each node or random variable (at time t) only depends on the nodes at previous time step (at time $t-1$). In essence, a DBN provides a suitable framework to represent uncertainties, dependencies and dynamics exhibited in the time series data; but its weakness remains in its time-invariant nature meaning that the underlying network structure remains unchanged over time: the dependency between inter-slice variables (connections across time slices) are fixed and invariant over time.
Let $\mathbf{X}^t = (X_1^t, \ldots, X_p^t)^\top \in \mathbb{R}^p$ be a vector representing the p random variables, at time t. A stochastic dynamic process can be modeled by a first-order Markovian process such that any variable at time t is dependent only on the variables observed at time $(t-1)$. One may restrict the transition network to contain only inter time slice interactions and avoid intra-slice interactions, which renders the dynamic Bayesian network as a first order multivariate time series process.
The autoregressive model $p(\mathbf{X}^t|\mathbf{X}^{t-1})$ defines a probabilistic distribution of random variables at time t given those at time $t-1$. The likelihood of the observed random variables over a time series of T steps can be expressed as:

$$p(\mathbf{X}, \ldots, \mathbf{X}^T) = p(\mathbf{X}^1) \prod_{t=2}^{T} p(\mathbf{X}^t|\mathbf{X}^{t-1}). \tag{2.63}$$

In this work, we assume that the transitional probability $p(\mathbf{X}^t|\mathbf{X}^{t-1})$ is represented by a linear model such that:

$$p(\mathbf{X}^t) \sim N(\mathbf{A}.\mathbf{X}^{t-1}, \sigma^2\mathbf{I}) \tag{2.64}$$

where $\mathbf{A} \in \mathbb{R}^{p\times p}$ is a matrix of coefficients and ϵ is a white noise with variance σ^2. The former offers a consistent estimation of the structure of DBNs, which can be recovered by reading off the nonzero coefficients $a_{ij} \in \mathbf{A}$ then adding an arc from $X_j^{(t-1)}$ to X_i^t respectively (Nagarajan et al., 2013).
One can search for a network structure using a scoring function, used to evaluate

how well the searched structure matches the data. The searching operator generally starts with a topology with no links, then iteratively searches for a possible structure by adding, reversing or deleting edges —there can be no cycles since all arrows go forward—so as to find a network that maximises the scoring function. This process is repeated until no modification could improve the score. There is a wide range of score-based methods that one could choose from to learn the DBN structure; for example: the likelihood score, the Minimum Description Length principle, the BIC/AIC score, the marginal likelihood and so forth.

2.6.1 Least Angle Regression

Least Angle Regression (LARS) is a model selection algorithm less greedy than conventional forward selection algorithms and that consists of fitting linear regression models to high-dimensional data (Efron et al., 2004). LARS is a fast computational algorithm, effective in high-dimensional settings, and can be easily modified to implement Lasso (Tibshirani, 1994). LARS starts with all coefficient equal to zero (like forward selection using Lasso) and finds the predictor most correlated with the response. The algorithm takes a step towards that predictor until finding another one, which has high correlation with the current residuals, that would deviate the trajectory into an equiangular direction (between the two predictors) until a third one appears and so forth. At each step LARS adds one covariate to the model; for example, if two variables are almost equally correlated with the response, then the coefficients of these variables increase at the same rate approximately. The sparsity of the network is enforced by introducing the lasso penalty (i.e. L_1-norm) and inferred based on the data using cross-validation technique. Dynamic Bayesian networks can be written nonparametrically as

$$X_i(t) = f(X_1(t-1), \ldots, \ldots, X_p(t-1)), \tag{2.65}$$

where t describes time index. If we assume a vector autoregressive process of the first order, then each variable X_i, $i = 1, \ldots, p$ can be represented as

$$X_i(t) = b_i + \sum_{j=1}^{p} a_{ij} X_j(t-1) + \epsilon_i, \tag{2.66}$$

where the nonzero elements a_{ij} are the adjacency matrix of the interaction network. The interpretation is that the variable j influences variable i if a_{ij} is not

zero. LARS tends to produce some coefficients exactly to zero by applying an L_1 norm penalty to their sum. Then, only nonzero coefficients define significant dependence relationships.

2.6.2 G1DBN

Lèbre (2009) proposed a two-step procedure for DBN inference: first, it learns a directed acyclic graph (DAG) encoding first-order conditional dependence $\mathcal{G}^{(1)}$ of each pair $(X_i(t), X_j(t-1))$ given all the rest at $t-1$. Linear dependencies (the partial regression coefficients $a_{ij|k}$, $k = 1, \ldots, p$, and p is the number of variables) can be defined as

$$X_i(t) = b_{ijk} + a_{ij|k} X_j(t-1) + a_{ik|j} X_k(t-1) + \epsilon_{ijk}(t), \tag{2.67}$$

where the rank of the matrix $(X_j(t-1), X_k(t-1))_{t\geq 2}$ equals to two and the noise is Gaussian. The conditional dependence between the variables $X_i(t)$ and $X_j(t-1)$ given other variables $X_k(t-1)$ is measured by testing the null assumption $\mathcal{H}_0^{i,j,k} : a_{ij|k} = 0$. For each $k \neq j$, the estimates $\hat{a}_{ij|k}$ are computed either by the: Least Square estimator, Huber estimator, or the Tukey bisquare estimator. This procedure generates the p-value $p_{ij|k}$ from the standard significance test:

$$\text{under} \quad (\mathcal{H}_0^{i,j,k}) : a_{ij|k} = 0, \qquad \frac{\hat{a}_{ij|k}}{\hat{\sigma}(\hat{a}_{ij|k})} \sim t(n-4), \tag{2.68}$$

where $t(n-4)$ refers to a Student probability distribution with $n-4$ degrees of freedom and $\hat{\sigma}(\hat{a}_{ij|k})$ is the variance estimates for $\hat{a}_{ij|k}$. A score $S_1(i,j)$ equal to $\max_{k\neq j}(p_{ij|k})$ of the $p-1$ (i.e. p is the number of variables) computed p-values derived from Equation(2.68) is being assigned for each potential edge $(X_j(t-1), X_i(t))$. This is the most favourable result to the first order conditional independence. This method does not derive p-values for the edges but allows to order the possible edges of DAG $\mathcal{G}^{(1)}$ according to how likely they are. The smaller the score or the p-value, the larger the significance of an edge becomes because in this case the null hypothesis that assumes $a_{ij|k} = 0$ is rejected. Eventually, weak edges are filtered out by selecting scores smaller or equal to an α_1 threshold, set by the user.

The second step consists of using the inferred DAG $\hat{\mathcal{G}}^{(1)}$ (from the previous step) to infer the real network structure $\tilde{\mathcal{G}}$. For each pair (i,j) such that the set of

edges $(X_j(t-1), X_i(t))$ is in $\hat{\mathcal{G}}^{(1)}$, the model is defined as,

$$X_i(t) = b_i + \sum_{j \in \text{pa}(X_i(t), \hat{\mathcal{G}}^{(1)})} a_{ij}^{(2)} X_j(t-1) + \epsilon_i(t), \tag{2.69}$$

where $a_{ij}^{(2)}$ is the regression coefficient and the rank of the matrix $(X_j(t-1))_{t \geq 2, j \in \text{pa}(X_i(t), \hat{\mathcal{G}}^{(1)})}$ is denoted by $|\text{pa}(X_i(t), \hat{\mathcal{G}}^{(1)})|$. Each edge of $\hat{\mathcal{G}}^{(1)}$ is assigned a score $S_2(i,j)$ equal to the p-value $p_{ij}^{(2)}$ derived from the significance test,

$$\text{under} \quad (\mathcal{H}_0^{i,j}) : a_{ij}^{(2)} = 0, \qquad \frac{\hat{a}_{ij}^{(2)}}{\hat{\sigma}(\hat{a}_{ij}^{(2)})} \sim t(n - 1 - |\text{pa}(X_i(t), \hat{\mathcal{G}}^{(1)})|). \tag{2.70}$$

The score $S_2(i,j) = 1$ is assigned to the edges that are not in $\hat{\mathcal{G}}^{(1)}$, the smallest scores indicate the most significant edges. Hence, the inferred DAG for $\tilde{\mathcal{G}}$ contains edges with a score below a specified threshold α_2.

2.6.3 Simone

SIMoNe (Statistical Inference for Modular Networks) (Chiquet et al., 2009) enables inference based on partial correlation coefficients with a Gaussian graphical model. The algorithm takes into account a latent network structure to increase the estimation accuracy, which assumes that each node belongs to some unobserved group. The latent clustering of the network is further used to drive the selection of arcs through an adaptive L_1 penalisation of the model likelihood. SIMoNe applies an Expectation maximisation (EM) strategy that alternates inference of the network latent structure and inference of the networks edges. In the E step, it estimates non-zero entries K_{ij} of the concentration matrix; then in the M step, it applies GLasso to infer the network edges (Friedman et al., 2007).

2.6.4 GeneNet

GeneNet is an R package for analyzing high-dimensional time series data (Opgen-Rhein and Strimmer, 2007) that estimates directed Gaussian graphical models (GGMs). Once the positive definitive covariance matrix is estimated (using a shrinkage estimator) as in (Schafer et al., 2006), it then computes the inverse

covariance matrix, also known as the concentration matrix, to derive the undirected GGM so that edges are related to the highest partial correlations or highest probabilistic dependence. GeneNet infers the directionality of the interactions by comparing, for each pair of connected nodes, the partial variances of the respective variables. An edge between two nodes is directed in such a fashion that the direction of the arrow points from the node with the larger standardized partial variance to the node with the smaller standardized partial variance. Each edge is given a p-value (the null hypothesis is that the partial correlation between its nodes is zero) and a score equal to 1 minus the respective p-value. Obviously the most effective edges are related with the lowest score.

2.7 Intractable Models

Bayesian inference is generally intractable for many models of practical interest because of the determination of the integration measure that is usually not feasible to be estimated in an analytical closed form, except for exponential families together with conjugate priors. In such situations, we need to resort to approximation schemes for achieving practical Bayesian inferences to successfully perform model comparison and prediction tasks. For this purpose, one can use either analytic approximations of integrals or methods based on Monte Carlo sampling: generally known as *approximate inference*. In this section I will review the *Sampling* approximation method for estimating the posterior distribution of intractable models.

2.7.1 Sampling approximation

In this section, I shall outline approximate methods based on numerical sampling algorithms that consist of generating random samples from a given distribution. These techniques are known as the *Monte Carlo* sampling. Suppose that we wish to evaluate the expectation of some function $g(x)$ with respect to a probability density function $f(x)$, given by

$$\mathbb{E}[g] = \int g(x)f(x)dx, \tag{2.71}$$

where x is a single continuous variable. The mean of the probability density function $f(x)$ can be obtained by taking $g(x) = x$, and the variance can be obtained by taking $g(x) = (x - \mu_x)^2$, as described in Equation (2.13) and (2.14) respectively. For simple $f(x)$ the integral expression described in Equation (2.71) can be evaluated analytically. However, this closed form solution becomes intractable when the target density function $f(x)$ increases in complexity. If we can simulate an independent set of samples $X = \{x^1, x^2, \ldots, x^L\}$ from the target density $f(x)$, we may then form the corresponding set of realisations denoted by $f(X) = \{f(x^1), f(x^2), \ldots, f(x^L)\}$, and hence for a large set of samples (when $L \to \infty$) the strong law of large numbers allows us to approximate the target density $f(x)$ by a finite sum

$$\hat{f} = \frac{1}{L} \sum_{l=1}^{L} \delta_{x^l}(dX), \tag{2.72}$$

where $\delta_{x^l}(dX)$ denotes the delta-Dirac mass located at x^l. The approximated target density $\hat{f}$ is obtained by finding the number of samples falling within the interval $[X, X + dX]$ divided by the total number of samples L. By plugging Equation (2.72) in (2.71), we obtain

$$\mathbb{E}[\hat{g}] = \int g(x)\hat{f}dx = \int g(x) \left(\frac{1}{L} \sum_{l=1}^{L} \delta_{x^l}(dX) \right) dx = \frac{1}{L} \sum_{l=1}^{L} g(x^l). \tag{2.73}$$

We deduce that in the limit of large samples the convergence to $\mathbb{E}[g]$ is almost sure, so that

$$\mathbb{E}[\hat{g}] = \frac{1}{L} \sum_{l=1}^{L} \mathbb{E}[g(x^l)] \xrightarrow[L\to\infty]{a.s.} \mathbb{E}[g]. \tag{2.74}$$

Moreover, the variance of the approximation can be written in the form

$$\text{var}[\hat{g}] = \frac{1}{L}\mathbb{E}[(g - \mathbb{E})^2]. \tag{2.75}$$

This method assumes that one could simulate an independent set of samples from a density distribution $f(x)$.

2.7.1.1 Rejection sampling

This method provides a basic sampling technique to generate observations from a complex density distribution $f(x)$. It consists of bounding the density distribution $f(x)$, from above, with a simple *proposal distribution* $q(x)$ from which we can readily draw samples. Then we weight each sample x^l generated from $q(x)$ in such a way to become a sample of the target distribution $f(x)$. We begin by introducing a constant M that satisfies $f(x) \leq Mq(x)$, such that $M < \infty$, to ensure that it dominates $f(x)$. Next, we evaluate $Mq(x^l)$ and then we multiply it by u, which is a randomly generated number from a uniform distribution $\mathcal{U}(0, 1)$. Finally, the result is compared to $f(x^l)$, if $uMq(x^l) \leq f(x^l)$ then the sample is accepted, otherwise it is rejected. The rejection sampling algorithm can be described as in Algorithm 2. Figure 2.9 illustrates the rejection sampling algorithm

Algorithm 2 Rejection Sampling algorithm

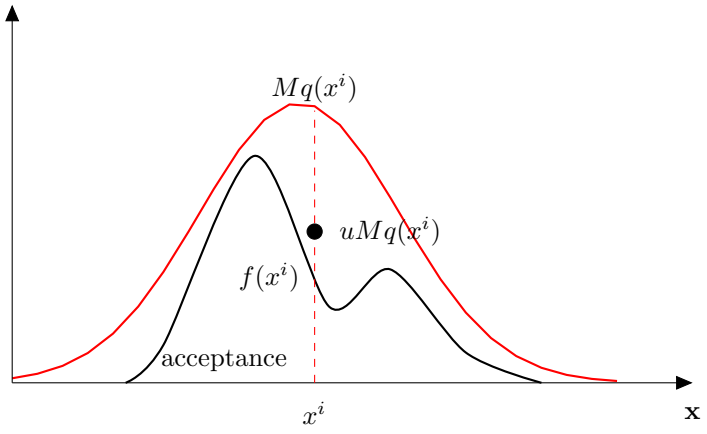
1: Draw L samples from $q(x)$, such that $X = \{x^1, \ldots, x^L\}$
2: **for** $l \leftarrow 1, L$ **do**
3: Generate a value u from a uniform distribution $\mathcal{U}(0, 1)$
4: **if** $u \leq \frac{f(x^l)}{Mq(x^l)}$ **then**
5: Accept the sample x^l
6: **else**
7: Reject the sample x^l
8: **end if**
9: **end for**
10: **return** a collection of samples $\{x^l\}$

where samples are drawn from a simple distribution $q(x)$, which is chosen to be Gaussian. If the generated sample falls within the distribution $f(x)$ we accept it, else we reject it. Therefore, after generating few thousand samples, we achieve approximating the density distribution $f(x)$. This is a basic sampling technique that suffers from several limitations

- If the proposed distribution $q(x)$ is very different from the target distribution $f(x)$, we may throw away a large number of samples before obtaining a good sample size of the target distribution.

- If we choose a large envelope to bound the target distribution from above, we end up rejecting a large number of samples. Hence, it becomes useful to select a tight envelope for defining the proposal distribution $q(x)$.

Figure 2.9: Illustration of the rejection sampling. The black curve represents the density function $f(x)$ and the red one represents the proposal distribution, which is a Gaussian. we sample a candidate x^i and a uniform variable u, then we accept the candidate sample if $uMq(x^i) < f(x^i)$, otherwise we reject it [adapted from (Andrieu et al., 2003)].



- This technique may be applicable for a univariate case, but it is impractical in high-dimensional situations, because the probability mass shifts away from the region of high probability density and becomes concentrated in a thin shell at large radius (or in the tails of the distribution) (Bishop, 2007, page 36). This is also known as *heavy tailed distributions* in statistics.

2.7.1.2 Importance sampling

As in the case of rejection sampling, importance sampling is based on an arbitrary proposal importance distribution $q(x)$ such that its support include the support of $f(x)$. Importance sampling performs better than rejection sampling because it does not throw away samples, but instead it employs them for approximating the expectation function using their importance weights, so that

$$\begin{aligned}\mathbb{E}[g] &= \int g(x)f(x)dx \\ &= \int g(x)\frac{f(x)}{q(x)}q(x)dx \\ &= \int g(x)w(x)q(x)dx,\end{aligned}$$

where w is known as the *importance weight* ratio, which represents the unbiased estimator, that is, used to rectify the distortion introduced by sampling from the

inappropriate distribution. It takes the value one when the sampled distribution $q(x)$ matches the distribution of $f(x)$. Consequently, if one can simulate L i.i.d. samples from $q(x)$, $X = \{x^l\}_{l=1}^L$, we can then write the Monte Carlo estimate in the form

$$\mathbb{E}[\hat{g}] \approx \frac{1}{L}\sum_{l=1}^{L} g(x^l)w(x^l). \tag{2.76}$$

This is known as the *importance sampling estimate*, which is valid if we can evaluate the importance weight $w(x)$ by which we can estimate the target distribution $f(x)$ in the form

$$\hat{f}(dX) = \frac{1}{L}\sum_{l=1}^{L} w(x^l)\delta_{x^l}(dX). \tag{2.77}$$

In Bayesian Statistics, the distribution of interest, $f(x)$, is evaluated up to a normalisation constant. We can then write $f(x) = \hat{f}(X)/\mathcal{Z}_f$, where $\hat{f}(X)$ is the unnormalised distribution that can be easily evaluated; whereas $\mathcal{Z}_f = \int \hat{f}(X)dX$ is an intractable normalisation constant. In this case we can write expectations as

$$\mathbb{E}[g] = \frac{\int g(X)\frac{\hat{f}(X)}{q(X)}q(X)dX}{\int \frac{\hat{f}(X)}{q(X)}q(X)dX} \tag{2.78}$$

By feeding Equation (2.78) with the approximated samples $\{x^l\}_{l=1}^L$ drawn from $q(x)$, we obtain

$$\begin{aligned}\mathbb{E}[\hat{g}] &\approx \frac{\sum_{l=1}^{L} g(x^l)\frac{\hat{f}(x^l)}{q(x^l)}}{\sum_{l=1}^{L}\frac{\hat{f}(x^l)}{q(x^l)}} \\ &= \sum_{l=1}^{L} g(x^l)\tilde{w}_l,\end{aligned} \tag{2.79}$$

where the *normalised importance weights* are defined by

$$\tilde{w}_l = \frac{\hat{f}(x^l)/q(x^l)}{\sum_{j=1}^{L}\hat{f}(x^j)/q(x^j)}. \tag{2.80}$$

If $\hat{f} = q$, then $\tilde{w}_l = 1/L$. Thus we can use the sample and weights $\{x^l, \tilde{w}^l\}_{l=1}^L$ to approximate any (suitable) expectation with respect to $\hat{f}$. Unfortunately, this method becomes inappropriate when the variability between $\tilde{f}$ and q increases because the weight vector will distort the non-zero components of the Equation

(2.79). This mismatch can be caused by the dimensionality of the samples X that could provoke the variability to grow in an exponential rate. An alternative solution is to use the *Sampling-importance-resampling* (SIR) method proposed by (Rubin, 1988), which is severely criticised by Andrieu et al. (2003) because of the fact that the SIR procedure does not provide a clear method for treating the high-dimensional problems. That is, because the resampling scheme introduces further Monte Carlo variation.

2.7.1.3 Markov Chain Monte Carlo

Markov chain Monte Carlo (MCMC) remains without doubt the most popular tool to compute approximate posterior inferences for Bayesian models. It can be applied to sample from any posterior distribution $p(\boldsymbol{\theta}|\mathbf{X})$ provided that we can evaluate analytically the expression presented in the numerator, as appears in Equation (2.50). The design matrix $\mathbf{X}$ comprises N samples each with D covariates. MCMC approximates the posterior distribution by a set of samples $\{\boldsymbol{\theta}^l\}_{l=1}^L$ rather than a closed form solution. Suppose that we need to generate a sample from an intractable posterior distribution $p(\boldsymbol{\theta}|\mathbf{X})$ for $\boldsymbol{\theta} \in \Theta$. We can achieve this by using a Markov chain with state space Θ, such that the set $\{\boldsymbol{\theta}^1, \boldsymbol{\theta}^2, \ldots, \boldsymbol{\theta}^l, \ldots, \boldsymbol{\theta}^L\}$ are realisations of the chain, which may be discrete or continuous. Here, I review some important properties of MCMC. A first order Markov chain is a stochastic process where the future state depends on the value of the present state, so that

$$p(\boldsymbol{\theta}^{l+1}|\boldsymbol{\theta}^1, \boldsymbol{\theta}^2, \ldots, \boldsymbol{\theta}^l) = p(\boldsymbol{\theta}^{l+1}|\boldsymbol{\theta}^l). \tag{2.81}$$

We say that $p(\boldsymbol{\theta}^{l+1}|\boldsymbol{\theta}^l)$ is the transition probability of the chain; a Markov chain is *homogeneous* if the transition probability is the same for all single steps, so that

$$\underbrace{p(\boldsymbol{\theta}^{l+1} = j|\boldsymbol{\theta}^l = i)}_{t_{i,j}(l)} = \underbrace{p(\boldsymbol{\theta}^{m+1} = j|\boldsymbol{\theta}^m = i)}_{t_{i,j}(m)} \qquad \forall l, m \in \mathbb{N}^+.$$

However, a *n-step* transition probability of a homogeneous Markov chain, can be defined by

$$t_{i,j}^{(n)} = p(\boldsymbol{\theta}^{l+n} = j|\boldsymbol{\theta}^l = i).$$

The *Chapman-Kolmogorov* equation is defined by using the Markov property and the law of total probability, so that

$$t_{i,j}^{(n+m)} = \sum_{k \in E} t_{i,k}^{(n)} t_{k,j}^{(m)} \quad \forall n, m \geq 0, \text{ and } i, j \in E, \tag{2.82}$$

where E is a countable set. This produces $(n+m)$*step* transition probability that evaluates the transition from the state i to j. A distribution $p(\boldsymbol{\theta})$ is stationary if it satisfies the marginal property, given by

$$p(\boldsymbol{\theta}^{l+1}) = \sum_{\boldsymbol{\theta}^{l}} (t_{l,l+1})\, p(\boldsymbol{\theta}^{l}). \tag{2.83}$$

A sufficient condition for ensuring a particular probability distribution $p(\boldsymbol{\theta}^l)$ to be invariant is to choose a transition probability to satisfy the *detailed balance equation*, so that

$$(t_{l,l+1})\, p(\boldsymbol{\theta}^{l}) = (t_{l+1,l})\, p(\boldsymbol{\theta}^{l+1}). \tag{2.84}$$

When the Markov chain satisfies the detailed balance equation, then the chain has $p(\boldsymbol{\theta})$ as a stationary distribution and so the reversed chain is homogeneous as well as *reversible*, such that

$$\begin{aligned} \sum_{\boldsymbol{\theta}^{l}} (t_{l,l+1})\, p(\boldsymbol{\theta}^{l}) &= \sum_{\boldsymbol{\theta}^{l}} (t_{l+1,l})\, p(\boldsymbol{\theta}^{l+1}) \\ &= p(\boldsymbol{\theta}^{l+1}) \sum_{\boldsymbol{\theta}^{l}} (t_{l+1,l}) \\ &= p(\boldsymbol{\theta}^{l+1}), \end{aligned} \tag{2.85}$$

Hence the obtained samples $\{\boldsymbol{\theta}^l\}_{l=1}^{L}$ are used to estimate the posterior distribution, as in Equation (2.74). It should also be noted that the early samples collected during the *burn in* stage must be discarded because we believe that those samples are only used for the sake of convergence.
Although this procedure defines a Markov chain which leaves the posterior invariant, it does not guarantee to visit any possible value under the posterior distribution. This requires two additional conditions to be fulfilled:

- The chain must be *irreducible*, which means if there exist a path with finite n-steps such that starting from any $\boldsymbol{\theta}^i$, there is a positive probability ($t_{i,j}^{(n)} > 0$) of visiting any other state $\boldsymbol{\theta}^j$.

- The chain must be *aperiodic*, which means that it does not get trapped into cycles.

A Markov chain satisfying these two conditions is termed *ergodic*, which converges to its unique equilibrium distribution regardless of the choice of the initial state. Several extensions to the MCMC method have been proposed; in what follows, I outline four approximation methods that have been commonly used for approximating the posterior distribution in Bayesian inferences.

2.7.1.4 Metropolis-Hastings sampling

The Metropolis-Hastings algorithm (Metropolis et al., 1953; Hastings, 1970) consists of constructing a successive set of samples $\{\boldsymbol{\theta}^1, \boldsymbol{\theta}^2, \ldots, \boldsymbol{\theta}^l, \ldots, \boldsymbol{\theta}^L\}$, with state space Θ, drawn from a Markov chain with stationary distribution $p(\boldsymbol{\theta}|\mathbf{X})$, that is, defined through a proposal distribution $q(\boldsymbol{\theta}^{l+1}|\boldsymbol{\theta}^l)$.
In general, if we define the distribution $p(\boldsymbol{\theta})$ up to a normalisation constant, we can then write

$$p(\boldsymbol{\theta}) = \frac{f(\boldsymbol{\theta})}{\mathcal{Z}},$$

where $\mathcal{Z}$ is an intractable normalisation constant. In this situation, it is not possible to sample directly from $p(\boldsymbol{\theta})$, but one may draw successive samples from a tractable arbitrary distribution, $q(\boldsymbol{\theta})$, which does not have to be a conditional version of $p(\boldsymbol{\theta})$. Let $\boldsymbol{\theta}^l$ be an initial sample drawn from an arbitrary distribution $q(\boldsymbol{\theta})$. In order to construct the transition probability we need to choose another sample $\boldsymbol{\theta}^*$ and compute the corresponding transition probability between these two states, $p(\boldsymbol{\theta}^*|\boldsymbol{\theta}^l)$. We may then accept or reject the new sample based on a probability measure $a(\boldsymbol{\theta}^*|\boldsymbol{\theta}^l)$, given by

$$a(\boldsymbol{\theta}^*|\boldsymbol{\theta}^l) = \min\left\{1, \frac{p(\boldsymbol{\theta}^*|\mathbf{X})q(\boldsymbol{\theta}^l|\boldsymbol{\theta}^*)}{p(\boldsymbol{\theta}^l|\mathbf{X})q(\boldsymbol{\theta}^*|\boldsymbol{\theta}^l)}\right\}. \tag{2.86}$$

The transition from one state to another is then based on the acceptance ratio a. For instance, if the proposal is accepted, we then accept $\boldsymbol{\theta}^{l+1} = \boldsymbol{\theta}^*$; otherwise, we reject the sample generated by $q(\boldsymbol{\theta}^*|\boldsymbol{\theta}^l)$ and set $\boldsymbol{\theta}^{l+1} = \boldsymbol{\theta}^l$ (as described in Algorithm 3).
We can see that the acceptance ratio depends on the posterior distribution for the initial and proposed states, $p(\boldsymbol{\theta}^*|\mathbf{X})/p(\boldsymbol{\theta}^l|\mathbf{X})$. This means that we can implement the acceptance ratio by multiplying the likelihood function by the prior

distribution, and disregard the intractable normalisation constant $\mathcal{Z}$. The main conditions that we need to ensure about the transition density is to hold the *irreducibility* and *aperiodicity* conditions on the state space. Under these conditions, the Markov chain will settle at the stationary distribution.

Algorithm 3 Metropolis-Hastings MCMC algorithm, adapted from (Barber, 2011)

Pick up the initial state, $\boldsymbol{\theta}^l$.
Propose a new sample $\boldsymbol{\theta}^*$ from the proposal $q(\boldsymbol{\theta}^*|\boldsymbol{\theta}^l)$.
Let $a = \frac{p(\boldsymbol{\theta}^*|\mathbf{X})q(\boldsymbol{\theta}^l|\boldsymbol{\theta}^*)}{p(\boldsymbol{\theta}^l|\mathbf{X})q(\boldsymbol{\theta}^*|\boldsymbol{\theta}^l)}$
if $a = 1$ **then**
    $\boldsymbol{\theta}^{l+1} = \boldsymbol{\theta}^*$ Acceptable proposal
else
    Generate a value u from a uniform distribution $U(u|0, 1)$.
    if $u \leq a$ **then**
        $\boldsymbol{\theta}^{l+1} = \boldsymbol{\theta}^*$ Acceptable proposal
    else
        $\boldsymbol{\theta}^{l+1} = \boldsymbol{\theta}^l$ Reject the proposal, and stay at $\boldsymbol{\theta}^l$ for an extra turn
    end if
end if
Repeat

2.7.1.5 Gibbs Sampler

The Gibbs sampler is a special case of the single components Metropolis-Hastings algorithm wherein the proposal sample is always accepted, such that $a = 1$. The key concept of the Gibbs algorithm is to sample sequentially from the posterior conditioned on the remaining parameters using a univariate conditional distributions. The advantage of using such methods is to make the computation fast and more tractable compared to other complex forms.

For example, suppose that we have three random variables $\boldsymbol{\theta}_1$, $\boldsymbol{\theta}_2$ and $\boldsymbol{\theta}_3$ and we wish to find the marginals $p(\boldsymbol{\theta}_1)$, $p(\boldsymbol{\theta}_2)$ and $p(\boldsymbol{\theta}_3)$; the sampler starts by defining initial values to these variables (i.e. $\boldsymbol{\theta}_1^{(0)}$, $\boldsymbol{\theta}_2^{(0)}$ and $\boldsymbol{\theta}_3^{(0)}$) and draws a sample for each component in an iterative way, such that

$$
\begin{aligned}
&\text{draw} \quad \boldsymbol{\theta}_1^{(1)} \quad \text{from} \quad p(\boldsymbol{\theta}_1|\mathbf{X}, \boldsymbol{\theta}_2^{(0)}, \boldsymbol{\theta}_3^{(0)}),\\
&\text{draw} \quad \boldsymbol{\theta}_2^{(1)} \quad \text{from} \quad p(\boldsymbol{\theta}_2|\mathbf{X}, \boldsymbol{\theta}_1^{(1)}, \boldsymbol{\theta}_3^{(0)}),\\
&\text{draw} \quad \boldsymbol{\theta}_3^{(1)} \quad \text{from} \quad p(\boldsymbol{\theta}_3|\mathbf{X}, \boldsymbol{\theta}_1^{(1)}, \boldsymbol{\theta}_2^{(1)}),\\
&\vdots
\end{aligned}
$$

$$
\begin{array}{llll}
\text{draw} & \boldsymbol{\theta}_1^{(l)} & \text{from} & p(\boldsymbol{\theta}_1|\mathbf{X}, \boldsymbol{\theta}_2^{(l-1)}, \boldsymbol{\theta}_3^{(l-1)}), \\
\text{draw} & \boldsymbol{\theta}_2^{(l)} & \text{from} & p(\boldsymbol{\theta}_2|\mathbf{X}, \boldsymbol{\theta}_1^{(l)}, \boldsymbol{\theta}_3^{(l-1)}), \\
\text{draw} & \boldsymbol{\theta}_3^{(l)} & \text{from} & p(\boldsymbol{\theta}_3|\mathbf{X}, \boldsymbol{\theta}_1^{(l)}, \boldsymbol{\theta}_2^{(l)}), \\
\vdots & & & \\
\text{and so forth.} & & &
\end{array}
$$

To achieve a successful sampling our conditional distributions must be tractable such that the sequence of transition probabilities for each sampled vector $\boldsymbol{\theta}^{(l)} = \left[\boldsymbol{\theta}_1^{(l)}, \boldsymbol{\theta}_2^{(l)}, \boldsymbol{\theta}_3^{(l)}\right]$ would take the form

$$
t_{\boldsymbol{\theta}^{(l)}, \boldsymbol{\theta}^{(l+1)}} = \prod_{i=1}^{3} p\left(\boldsymbol{\theta}_i^{(l+1)} | \left(\boldsymbol{\theta}_{j/i}^{(l)} \text{ if } j > i\right) \text{ or } \left(\boldsymbol{\theta}_{j/i}^{(l+1)} \text{ if } j < i\right), \mathbf{X}\right) \quad \text{where } j \in \{1, 2, 3\}. \tag{2.87}
$$

For an infinitely large of sample size ($l \to \infty$), the Gibbs sequence converges to a stationary distribution, hence the sampled vector $\boldsymbol{\theta}^{(l)} = \left[\boldsymbol{\theta}_1^{(l)}, \boldsymbol{\theta}_2^{(l)}, \boldsymbol{\theta}_3^{(l)}\right]$ tends to be sampled from the true posterior. Accordingly, if we choose m replicates $\left(\boldsymbol{\theta}_1^{(l+1)}, \ldots, \boldsymbol{\theta}_1^{(l+m)}\right)$ after discarding sufficient burn-in samples, we may then obtain the marginals for each distribution respectively by approximating the samples as derived from $p(\boldsymbol{\theta}_i|\mathbf{X})$ where $i \in \{1, 2, 3\}$.
The main advantage of Gibbs sampling is that it saves us from choosing a proposal density function, but on the other hand it suffers from the fact that it may not converge for highly correlated variables.

2.7.1.6 Metropolis Adjusted Langevin Algorithm

Stramer and Tweedie (1999) developed a proposal distribution based on Langevin-type diffusions that converges faster than traditional random walk methods usually suffering from slow convergence and long runtimes. Consider the parameter vector $\boldsymbol{\theta} \in \mathbb{R}^D$ with density $p(\boldsymbol{\theta})$. For example by setting $\boldsymbol{\theta} = \{\theta_1, \theta_2, \theta_3, \theta_4\} \in \mathbb{R}^4$, $D = 4$, the Metropolis adjusted Langevin algorithm (MALA) relies on the gradient of the log-density $\mathcal{L}(\boldsymbol{\theta}) \equiv \log(p(\boldsymbol{\theta}))$ to make proposals satisfying the Langevin diffusion defined by the stochastical differential equation, such that

$$
\mathrm{d}\boldsymbol{\theta}(t) = \nabla_{\boldsymbol{\theta}} \mathcal{L}\{\boldsymbol{\theta}(t)\}\mathrm{d}t/2 + \mathrm{d}\boldsymbol{b}(t) \tag{2.88}
$$

where $\boldsymbol{b}$ is a 4-dimensional Brownian motion. In practice, one cannot simulate directly from Equation (2.88), but instead it is recommended to use a discretised

form of the dynamic. A common choice for solving this discretisation problem is to use the first order Euler approximation that can be defined as

$$\boldsymbol{\theta}^* = \boldsymbol{\theta}^n + \frac{\epsilon^2}{2}\nabla_{\boldsymbol{\theta}}\mathcal{L}\{\boldsymbol{\theta}^n\} + \epsilon \boldsymbol{z}^n \tag{2.89}$$

where $\epsilon > 0$ is the integration step size, and $\boldsymbol{z}$ are normally distributed $\mathcal{N}(\mathbf{0}, \boldsymbol{I})$ independent random variables. The proposal density would then become

$$q(\boldsymbol{\theta}^*|\boldsymbol{\theta}^n) = \mathcal{N}(\boldsymbol{\theta}^*|\boldsymbol{\theta}^n + \frac{\epsilon^2}{2}\nabla_{\boldsymbol{\theta}}\mathcal{L}\{\boldsymbol{\theta}^n\}, \epsilon^2\boldsymbol{I}). \tag{2.90}$$

This naive first order Euler discretisation suffers from errors due to discretisation that affects the rate of convergence (Stramer and Tweedie, 1999). The discretisation errors are reduced by employing the Metropolis acceptance probability after each integration step. Roberts and Rosenthal (1998) analysed the optimal scaling of ϵ to Langevin diffusions as $D \to \infty$. One can increase the acceptance rate when scaling ϵ by $D^{1/3}$, as cited in Roberts and Rosenthal (1998). Thus MALA requires $\mathcal{O}(D^{1/3})$ steps to converge.

2.7.1.7 Hamiltonian Monte Carlo

The Hybrid Monte Carlo sampling technique (Duane et al., 1987) found its origin in statistical physics literature as a MCMC technique for sampling from a complex physical system. This technique was applied to statistical inference problems (Neal, 1992, 1993a,b) to obtain samples from the posterior distribution for Bayesian neural networks. It is also known as Hamiltonian Monte Carlo (HMC) because it employs Hamiltonian dynamics between states $\boldsymbol{\theta} \in \mathbb{R}^D$ and augmented variables $\boldsymbol{p} \in \mathbb{R}^D$ so as to move from one point $\boldsymbol{\theta}, \boldsymbol{p}$ to another $\boldsymbol{\theta}^*, \boldsymbol{p}^*$ located further away in space and which will be accepted with a high probability. The density of $\boldsymbol{\theta}$ is described as $p(\boldsymbol{\theta})$, and the density of the auxiliary variable of $\boldsymbol{p}$ is described as $p(\boldsymbol{p}) = \mathcal{N}(\boldsymbol{p}|\mathbf{0}, \boldsymbol{M})$ where $\boldsymbol{M}$ is the covariance matrix. The auxiliary variables have Gaussian distributions, independent of $\boldsymbol{\theta}$, and of each other. The factorised joint distribution can be described as

$$\begin{aligned} p(\boldsymbol{\theta}, \boldsymbol{p}) &= p(\boldsymbol{\theta})p(\boldsymbol{p}) \\ &= \frac{1}{Z_{\boldsymbol{\theta}}} e^{H_{\boldsymbol{\theta}}(\boldsymbol{\theta})} \frac{1}{Z_{\boldsymbol{p}}} e^{H_{\boldsymbol{p}}(\boldsymbol{p})} = \frac{1}{Z} e^{(H_{\boldsymbol{\theta}}(\boldsymbol{\theta}) + H_{\boldsymbol{p}}(\boldsymbol{p}))} = \frac{1}{Z} e^{H(\boldsymbol{\theta}, \boldsymbol{p})}. \end{aligned} \tag{2.91}$$

The Hamiltonian can be defined as

$$H(\boldsymbol{\theta}, \boldsymbol{p}) = \underbrace{-\mathcal{L}(\boldsymbol{\theta}) + \frac{1}{2}\log\{(2\pi)^D|\boldsymbol{M}|\}}_{\text{potential energy}} + \underbrace{\frac{1}{2}\boldsymbol{p}^\top \boldsymbol{M}^{-1}\boldsymbol{p}}_{\text{kinetic energy}} , \tag{2.92}$$

where $\mathcal{L}(\boldsymbol{\theta}) \equiv \log(p(\boldsymbol{\theta}))$, $H(\boldsymbol{\theta}, \boldsymbol{p})$ can be viewed as the sum of a 'potential energy' with a 'kinetic energy', and $|M|$ represents the mass matrix.
The derivatives of $\boldsymbol{\theta}$ and $\boldsymbol{p}$ with respect to a fictitious time variable τ can be described as

$$\begin{aligned} \frac{d\boldsymbol{\theta}}{d\tau} &= \frac{\partial H}{\partial \boldsymbol{p}} = \boldsymbol{M}^{-1}\boldsymbol{p} \\ \frac{d\boldsymbol{p}}{d\tau} &= -\frac{\partial H}{\partial \boldsymbol{\theta}} = \nabla_{\boldsymbol{\theta}}\mathcal{L}(\boldsymbol{\theta}). \end{aligned} \tag{2.93}$$

This differential equation cannot be solved analytically, instead one needs to recourse to numerical approximation technique. A common choice for this problem consists of employing the Stromer-Verlet or 'leapfrog' integrator as used in (Duane et al., 1987; Neal, 1992, 1993b,a), which consists of applying the following steps:

$$\boldsymbol{p}(\tau + \epsilon/2) = \boldsymbol{p}(\tau) + \frac{\epsilon}{2}\nabla_{\boldsymbol{\theta}}\mathcal{L}\{\boldsymbol{\theta}(\tau)\}, \tag{2.94}$$

$$\boldsymbol{\theta}(\tau + \epsilon) = \boldsymbol{\theta}(\tau) + \epsilon\boldsymbol{M}^{-1}\boldsymbol{p}(\tau + \epsilon/2), \tag{2.95}$$

$$\boldsymbol{p}(\tau + \epsilon) = \boldsymbol{p}(\tau + \epsilon/2) + \frac{\epsilon}{2}\nabla_{\boldsymbol{\theta}}\mathcal{L}\{\boldsymbol{\theta}(\tau + \epsilon)\}. \tag{2.96}$$

The time reversibility can be easily verified by changing the sign of the step size ϵ. The integrator preserves phase space volume by showing that the Jacobian transformation from $\boldsymbol{\theta}, \boldsymbol{p}$ at time τ to $\boldsymbol{\theta}, \boldsymbol{p}$ at time $\tau + \epsilon$ has a unit determinant (Neal, 1992). Equation(2.94) to (2.96) can be iterated several times (e.g. $L = 50$) to generate a candidate state $\boldsymbol{\theta}^*, \boldsymbol{p}^*$. This proposal step is then accepted if $H(\boldsymbol{\theta}^*, \boldsymbol{p}^*) < H(\boldsymbol{\theta}, \boldsymbol{p})$, otherwise it is accepted with a probability $\min\{1, \exp\{-H(\boldsymbol{\theta}^*, \boldsymbol{p}^*) + H(\boldsymbol{\theta}, \boldsymbol{p})\}\}$. A simplified form of the model can be obtained by using $\boldsymbol{M} = \boldsymbol{I}$, identity matrix. The HMC algorithm that I used in chapter 3 to sample from the posterior distributions can be described as in Algorithm 4.

Algorithm 4 HMC algorithm

for $chain \leftarrow 1, n.chain$ **do**
  Initialise the set of parameters $\boldsymbol{\theta} = \{\theta_1, \theta_2, \theta_3, \theta_4\}$.
  Set the gradient $\boldsymbol{g} = Grad(\boldsymbol{\theta})$, and the log-posterior probability such that $p(\boldsymbol{\theta}) = \mathcal{L}(\boldsymbol{\theta}) + \log(\pi(\boldsymbol{\theta}))$.
  for $iter \leftarrow 1, n.iter$ **do**
    Initialise the momentum '$\boldsymbol{P}$' using a normal distribution.
    Evaluate the Hamiltonian value $H = \text{Hamiltonian}(p(\boldsymbol{\theta}), \boldsymbol{P})$.
    Propose a new step $\boldsymbol{\theta}^*$ and find related gradient $\boldsymbol{g}^* = Grad(\boldsymbol{\theta}^*)$.
    for $l \leftarrow 1, L$ **do**
      $\boldsymbol{P} = \boldsymbol{P} + (\boldsymbol{\epsilon}/2) * \boldsymbol{g}^*$
      $\boldsymbol{\theta}^* = \boldsymbol{\theta}^* + \boldsymbol{\epsilon} * \boldsymbol{P}$
      $\boldsymbol{g}^* = Grad(\boldsymbol{\theta}^*)$
      $\boldsymbol{P} = \boldsymbol{P} + (\boldsymbol{\epsilon}/2) * \boldsymbol{g}^*$
    end for
    Evaluate $p(\boldsymbol{\theta}^*)$ and H^* values over $\boldsymbol{\theta}^*$.
    Accept or reject the state at end of trajectory, returning either the position at the end of the trajectory $\boldsymbol{\theta}^*$ or the initial position $\boldsymbol{\theta}$.
    $dH = H^* - H$.
    Accept the new proposal step in Monte-Carlo fashion.
    if $dH \leq 0$ **then**
      Acceptance steps, $\boldsymbol{\theta} = \boldsymbol{\theta}^*$.
    else
      if $dH \geq \mathcal{U}(1)$ **then**
        Acceptance steps, $\boldsymbol{\theta} = \boldsymbol{\theta}^*$.
      else
        Rejectance steps.
      end if
    end if
  end for
end for

2.8 Dynamical Models

2.8.1 State Space Models

The general setup for a state space model (SSM) consists of combining a state process $\{X_t\}_{t\geq 1}$ with an observation process $\{Y_t = f_t(X_t, \epsilon_t)\}_{t\geq 1}$, where $\{\epsilon_t\}$ is a sequence of independent random variables. The state process is an abstract quantity of the model that evolves in time and capable of generating an observation according to a given state —it constitutes a sufficient statistic for the model. The assumptions of a SSM can be defined as

i : The state sequence $\{X_t\}_{t\geq 0}$ is a first order Markov process, where the probability law of its process is identified by the initial density $p_0(X_0)$ and the transition densities $p(X_t|X_{t-1})$.

ii : The observations $\{Y_t\}_{t\geq 1}$ are conditionally independent of X_s and Y_s for $s \neq t$ given X_t.

Figure 2.10: Graphical representation of the SSM.

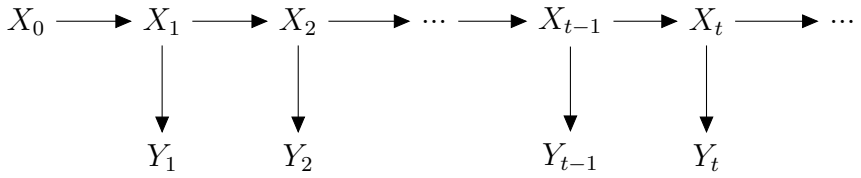


Figure 2.10 illustrates a general structure of a SSM representing the relationship between the state and observable processes such that the joint probability can be written in the form

$$p(x_{1:t}, y_{1:t}) = p(x_1)\left(\prod_{i=2}^{t} p(x_i|x_{i-1})\right)\left(\prod_{i=1}^{t} p(y_i|x_i)\right), \tag{2.97}$$

where $x_{1:t} = (x_1, \ldots, x_t)$ and $y_{1:t} = (y_1, \ldots, y_t)$ denote collections of states and observations ranging from time step 1 to t respectively. A primary concern in many state space inference problems is sequential estimation of the filtering distribution $p(x_t|y_{1:t})$ (Godsill et al., 2004). For a Linear-Gaussian model we can

write the likelihood and filtering distributions in the following form

$$p(y_t|y_{t-1}) = \mathcal{N}(y_t|\mathbf{A}y_{t-1}, \Gamma) \tag{2.98}$$

$$p(x_t|y_t) = \mathcal{N}(x_t|\mathbf{C}y_t, \Sigma), \tag{2.99}$$

where $\{\mathbf{A}, \Gamma, \mathbf{C}, \Sigma\}$ are the model parameters, which can be obtained by computationally efficient recursive formulas such as the Kalman filter (Kalman, 1960); but for general HMMs where the model is non-linear and the conditional distributions are non-Gaussian, the filtering posterior probability distribution of the state conditional on the observations can be written as

$$p(x_t|y_{1:t}) = \frac{p(x_t|y_{1:t-1})p(y_t|x_t)}{p(y_t|y_{1:t-1})} \propto p(y_t|x_t) \int p(x_t|x_{t-1})p(x_{t-1}|y_{1:t-1})dx_{t-1}, \tag{2.100}$$

where $t > 1$ and the posterior distribution for determining the initial state is given by

$$p(x_1|y_1) = \frac{p(x_1)p(y_1|x_1)}{p(y_1)} \propto p(x_1)p(y_1|x_1).$$

However, the likelihood can be described as

$$p(y_t|y_{1:t-1}) = \int \left\{ p(y_t|x_t) \int p(x_t|x_{t-1})p(x_{t-1}|y_{1:t-1})dx_{t-1} \right\} dx_t. \tag{2.101}$$

Therefore, numerical approximations are often required to solve the filtering equation.

2.8.2 Sequential Monte Carlo

Sequential Monte Carlo (SMC) methods first appeared under the name of the bootstrap filter (Gordon et al., 1993) inspired from the sampling importance resampling (SIR) method (Rubin, 1988), which aims to recursively evaluate the filter and the prediction densities by a set of randomly generated particles at each time step separately, such that

1. **Resample** an existing particle x_{t-1}^l with a probability $\tilde{w}_{t-1}^l$.

2. **Propagate** the particle to time t by sampling from $p(x_t|x_{t-1}^l)$.

We begin by assuming that the *particle filter*, a simulation-based filter, has fixed model parameters $\boldsymbol{\theta}$. Given the weighted particles $\{x_{t-1}^l, \tilde{w}_{t-1}^l\}_{l=1}^L$, we approximate Equation (2.100) by a weighted sample of particles $\{x_t^l, \tilde{w}_t^l\}_{l=1}^L$, such that

$$p(x_t|y_{1:t}) \propto p(y_t|x_t) \int p(x_t|x_{t-1}) p(x_{t-1}|y_{1:t-1}) dx_{t-1}.$$

Given the weighted particles $\{x_{t-1}^l, \tilde{w}_{t-1}^l\}_{l=1}^L$, we can write

$$\begin{aligned} \hat{p}(x_t|y_{1:t}) &\propto p(y_t|x_t) \sum_{l=1}^{L} \tilde{w}_{t-1}^l p(x_t|x_{t-1}^l) \\ &\propto \text{weight} \times \text{proposal}, \end{aligned}$$

which can be approximated using importance sampling. Accordingly, the posterior would be approximated with a set of weighted particles $\{x_t^l, \tilde{w}_t^l\}_{l=1}^L$. Whereas the prediction stage combines the current filtering distribution with the state evolution to obtain the prior distribution of the state at time $(t+1)$ via the Chapman-Kolmogorov equation

$$p(x_{t+1}|y_{1:t}) = \int p(x_{t+1}|x_t) p(x_t|y_{1:t}) dx_t. \tag{2.102}$$

The state equation can be defined as being used to generate the weight and sample values of the current state vector x_{t+1} based on past sampled values of the state x_t, which are typically not available in closed form for general nonlinear and non Gaussian state space models. A generic algorithm for the SMC method can be described as follows

1. At time step t, we have a sample representation of the conditional density of the state process $p(x_t|y_{1:t})$ expressed as samples $\{x_t^l\}_{l=1}^L$ with corresponding normalised importance weights $\{\tilde{w}^l\}_{l=1}^L$.

2. Resample with replacement the weighted samples $\{x_t^l\}_{l=1}^L$, according to their importance weights $\{\tilde{w}^l\}_{l=1}^L$, to obtain a set $\{\tilde{x}_t^l\}_{l=1}^L$. This has the role of eliminating particles with low importance weights.

3. Propagate each particle forward and approximate $\{x_{t+1}^l\}$ samples from the state transition density $p(x_{t+1}|\tilde{x}_t^l)$, giving approximate samples from $p(x_{t+1}|y_{1:t})$. Each of these particles are then perturbed slightly to perform

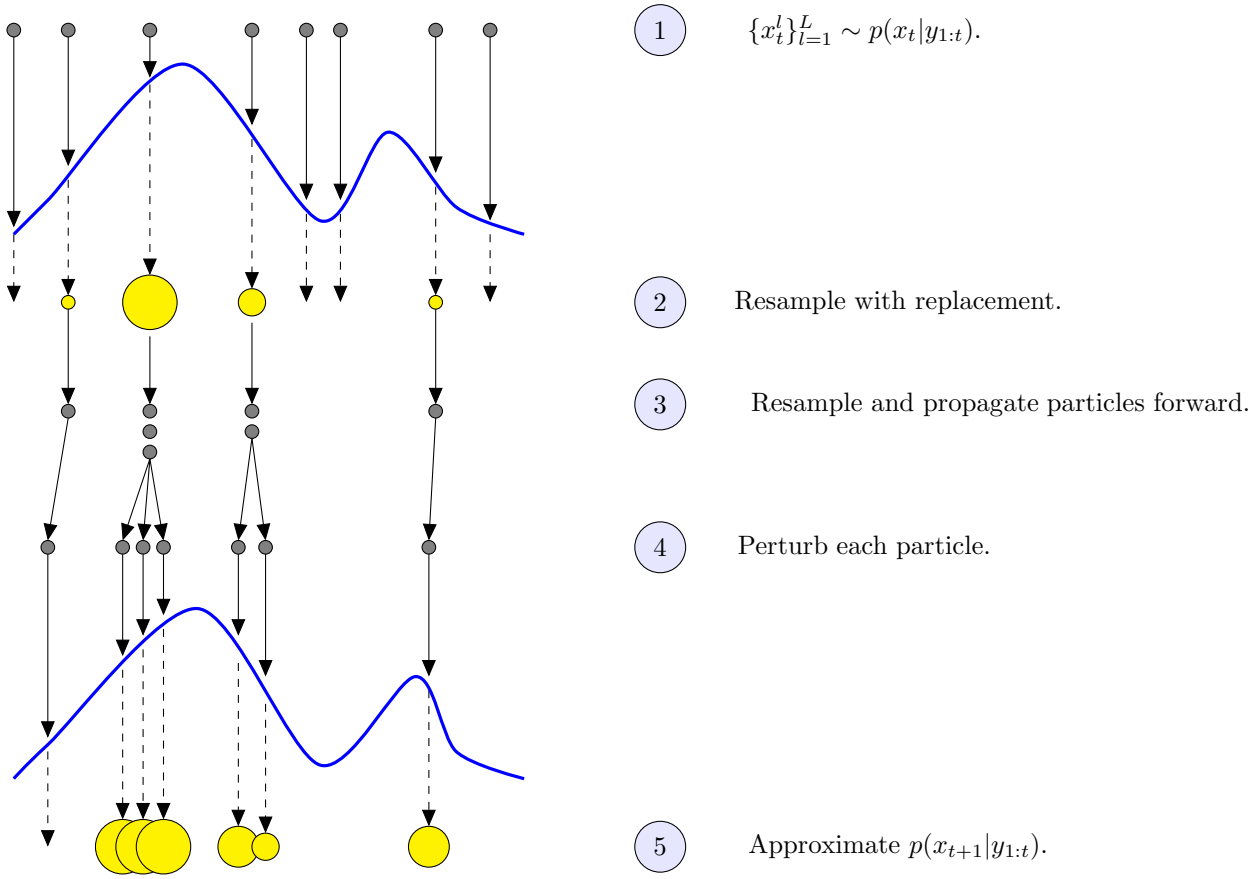


Figure 2.11: A graphical representation of the time evolution of a sequential Monte Carlo algorithm [adapted from (Doucet et al., 2001)]. At the first level, we generate $\{x_t^l\}_{l=1}^L$ samples from $p(x_t|y_{1:t})$. At the second level, we resample with replacement to clear away particles that fall below a certain threshold, known as low importance weight. At the third level, we resample the obtained particles according to their weights such that the heavier particles can be resampled more than once and then we propagate them forward to approximate $p(x_{t+1}|y_{1:t})$. At the fourth level, we perturb the particles in order to explore the state space better. Finally, we evaluate and normalise the importance weights of particles approximated from $p(x_{t+1}|y_{1:t})$. It should also be noted that the solid curve line describes the likelihood function at a certain time.

moves around the parameter space for better exploration.

4. For each of the new particles, evaluate the importance weights $w_{t+1} = p(y_{t+1}|x^l_{t+1})$, and then a normalised weight $\tilde{w}^l_{t+1} = w_{t+1}/\sum_i w^i_{t+1}$.

5. Go back to step 1 and repeat for $t = t + 1$.

Figure 2.11 shows a pictorial representation of the sequential Monte Carlo algorithm described above. Doucet et al. (2000) discussed that in the presence of lengthy observations the variance of the importance weights may increase over time, which makes the degeneracy phenomenon unavoidable. One way to limit this drawback is to choose an importance density that minimises the variance of the weights, known as the optimal importance density.

2.9 Marginal Likelihood Estimation

Gelman and Meng (1998) introduced a new technique known as path sampling for approximating the marginal likelihood of a model with parameter vector $\boldsymbol{\theta}$. Consider introducing an auxiliary variable (or temperature schedule) t ranging from 0 to 1, and a power posterior defined by various levels of weighted likelihood, such that

$$p_t(\boldsymbol{\theta}|\mathbf{X}) \propto p(\mathbf{X}|\boldsymbol{\theta})^t p(\boldsymbol{\theta}). \tag{2.103}$$

The auxiliary variable t is inspired by ideas from thermodynamics defined such that $t_0 = 0$ and $t_T = 1$. By choosing a temperature schedule that links t_0 to t_T, such that

$$0 = t_0 < t_1 < \cdots < t_T = 1, \tag{2.104}$$

we obtain a path between p_{t_0} and p_{t_T}: forming a natural bridge or path from the prior to the posterior distributions. The Bridge sampling technique improves the convergence of MCMC methods as it flattens the likelihood function for small values of t, prior-like distribution, and recovers the posterior of interest for $t = 1$. The posterior expectation estimator is given by

$$z(\mathbf{X}|t) = \int [p(\mathbf{X}|\boldsymbol{\theta})]^t p(\boldsymbol{\theta}) d\boldsymbol{\theta}, \tag{2.105}$$

so that $z(\mathbf{X}|t=0)$ is the integral of the prior, which is 1, while $z(\mathbf{X}|t=1)$ is the marginal likelihood $p(\mathbf{X}) = \int p(\mathbf{X}|\boldsymbol{\theta})p(\boldsymbol{\theta})d\boldsymbol{\theta}$, which can be written in a telescoping fashion as

$$p(\mathbf{X}) = z(\mathbf{X}|t=1) = z(\mathbf{X}|t_T) = \frac{z(\mathbf{X}|t_n)}{z(\mathbf{X}|t_0)} = \frac{z(\mathbf{X}|t_1)}{z(\mathbf{X}|t_0)} \times \frac{z(\mathbf{X}|t_2)}{z(\mathbf{X}|t_1)} \times \ldots \times \frac{z(\mathbf{X}|t_T)}{z(\mathbf{X}|t_{T-1})}. \tag{2.106}$$

Taking logarithms of both sides of Equation (2.106), we can derive an estimate of $z(\mathbf{X}|t=1)$ such that

$$\log\{p(\mathbf{X})\} = \log\left\{\frac{z(\mathbf{X}|t=1)}{z(\mathbf{X}|t=0)}\right\} = \int_0^1 E_{\boldsymbol{\theta}|\mathbf{X},t} \log\{p(\mathbf{X}|\boldsymbol{\theta})\}dt. \tag{2.107}$$

Thus the log marginal likelihood is the expected log likelihood with respect to the power posterior at temperature $t \in [0,1]$. Friel and Pettitt (2008) derived Equation (2.107) such as:

$$\begin{aligned}
\frac{d}{dt}\log\{z(\mathbf{X}|t)\} &= \frac{1}{z(\mathbf{X}|t)}\frac{d}{dt}z(\mathbf{X}|t) \\
&= \frac{1}{z(\mathbf{X}|t)}\frac{d}{dt}\int_{\boldsymbol{\theta}}[p(\mathbf{X}|\boldsymbol{\theta})]^t p(\boldsymbol{\theta})d\boldsymbol{\theta} \\
&= \frac{1}{z(\mathbf{X}|t)}\int_{\boldsymbol{\theta}}[p(\mathbf{X}|\boldsymbol{\theta})]^t \log\{p(\mathbf{X}|\boldsymbol{\theta})\}p(\boldsymbol{\theta})d\boldsymbol{\theta} \\
&= \int_{\boldsymbol{\theta}}\frac{[p(\mathbf{X}|\boldsymbol{\theta})]^t p(\boldsymbol{\theta})}{z(\mathbf{X}|t)}\log\{p(\mathbf{X}|\boldsymbol{\theta})\}d\boldsymbol{\theta} \\
&= E_{\boldsymbol{\theta}|\mathbf{X},t}[\log\{p(\mathbf{X}|\boldsymbol{\theta})\}].
\end{aligned}$$

Equation (2.107) is estimated by marginalising over the power parameter t and the model parameter vector $\boldsymbol{\theta}$. The method consists of: discretising the integral (2.107) over $t \in [0,1]$; run separate chains for each t; and sampling from the power posterior to estimate the expected log likelihood, $E_{\boldsymbol{\theta}|\mathbf{X},t}[\log\{p(\mathbf{X}|\boldsymbol{\theta})\}]$. The marginal likelihood can be evaluated using the trapezoid rule over T intervals defined using a temperature schedule of type, $t_s = a_s^c$, where $a_s = s/T$ is an equal spacing of T cutpoints in the interval $[0,1]$, and $c > 1$ is a constant that ensures the t_s are sampled with a high frequency in the region close to $t = 0$. Thus we

can approximate the logarithm of the evidence as

$$\log\{p(\mathbf{X})\} \approx \sum_{s=0}^{T-1}(t_{s+1}-t_s)\frac{1}{2}\left[E_{\boldsymbol{\theta}|\mathbf{X},t_{s+1}}[\log\{p(\mathbf{X}|\boldsymbol{\theta})\}] + E_{\boldsymbol{\theta}|\mathbf{X},t_s}[\log\{p(\mathbf{X}|\boldsymbol{\theta})\}]\right]. \tag{2.108}$$

The Monte Carlo standard error of $\log\{p(\mathbf{X})\}$ is obtained as the square root of the summed of variances at each cutpoint, $\nu_s = E_{\boldsymbol{\theta}|\mathbf{X},t_{s+1}}[\log\{p(\mathbf{X}|\boldsymbol{\theta})\}]$, so that,

$$\log\{p(\mathbf{X})\} \approx \sqrt{\left\{\frac{(t_2-t_1)^2}{2}\nu_1^2 + \sum_{s=2}^{T-1}\frac{(t_{s+1}-t_{s-1})^2}{2}\nu_s^2 + \frac{(t_T-t_{T-1})^2}{2}\nu_T^2\right\}}. \tag{2.109}$$

2.10 Random Processes

A random process $\mathcal{F}$, also called a stochastic process, is a family of random variables $\{f(x); \forall x \in \mathcal{X}\}$ that maps the sample space Ω into some set S. The choice of $\mathcal{X}$ and S define the characteristics of the process on whether it is countable or uncountable. For example, the Markov jump process is a 'discrete-time' process because the random variables are indexed by some set $\mathcal{X} = \{0, 1, 2, \ldots\}$; however, the Gaussian Process (see below) is a 'continuous-time' process because it is defined over a continuous set such that $\mathcal{X} = [0, \infty[$. Many important processes have the property that their joint probability distribution does not change under time shifts, which are known as the *stationary distributions*. In this section, I shall describe the Gaussian processes (GPs) and time series data analysis.

2.10.1 Gaussian processes

A stochastic process $\mathcal{F} = \{f(x); \forall x \in \mathcal{X}\}$ is said to be a *Gaussian process* (GP) if for any finite subset $Q \subset \mathcal{X}$, the obtained random vector $\mathcal{F} = \{f(x); \forall x \in Q\}$ is a multivariate Gaussian distribution. This can be interpreted as a distribution over function spaces.

In the context of machine learning, GPs belong to a non-parametric Bayesian framework which is defined over an infinite-dimensional parameter space. This approach affords to have models with infinitely many parameters because it regards the learning as an inference over the parameters of the model rather than an optimisation approach. The idea of replacing supervised neural networks by

Gaussian processes was first explored by Williams and Rasmussen (1996) and Neal (1997). In general, a GP can be described using the mean function $m(x)$ and a covariance or kernel function $k(x, x')$ of a continuous process $f(x)$, such that

$$\begin{aligned} m(x) &= \mathbb{E}[f(x)] \\ k(x, x') &= \mathbb{E}[(f(x) - m(x))(f(x') - m(x'))], \end{aligned} \tag{2.110}$$

where x and $x' \in \mathcal{X}$. Hence we will write the Gaussian process as

$$f(x) \sim \mathcal{GP}(m(x), k(x, x')). \tag{2.111}$$

To achieve a Bayesian learning approach, we need to specify a prior over the parameters with the purpose of expressing our beliefs about the parameters before we observe the data. Lawrence et al. (2010) have addressed the problem of how one may place a prior distribution over an infinite-dimensional object. For a finite number of inputs, $X = \{x_i\}_{i=1}^n$, the prior distribution over the vector $\mathbf{f} = [f(x_1), f(x_2), \ldots, f(x_n)]$ follows a Gaussian distribution $\mathcal{N}(\boldsymbol{\mu}, \mathbf{K}_{f,f})$, where $\boldsymbol{\mu}$ is the mean function and $\mathbf{K}_{f,f}$ is the covariance function obtained by evaluating the kernel function on the observed inputs. Hence, the prior distribution of the training set can be described in the form

$$p(\mathbf{f}) = \mathcal{N}(\mathbf{0}, \mathbf{K}_{f,f}), \tag{2.112}$$

which is a n-dimensional Gaussian distribution where $\mathbf{0}$ denotes a zero mean Gaussian prior and $\mathbf{K}_{f,f}$ is the $n \times n$ covariance matrix.
In practice, we observe noisy realisations of the continuous process, such that

$$y(x_i) = f(x_i) + \epsilon(x_i), \qquad \epsilon(x_i) \sim \mathcal{N}(0, \sigma_i^2) \qquad \forall i = \{1, 2, \ldots, n\}.$$

This can be used to define a Gaussian likelihood model such that

$$p(Y|\mathbf{f}, \sigma^2 I) = \mathcal{N}(\mathbf{f}, \sigma^2 I). \tag{2.113}$$

One may then define the marginal likelihood in the form

$$p(Y|X) = \int p(Y|\mathbf{f}, X)p(\mathbf{f}|X)d\mathbf{f} \\ = \mathcal{N}(Y|\mathbf{0}, \mathbf{K}_{f,f} + \sigma^2 I) \tag{2.114}$$

The logarithm of the marginal likelihood for the Gaussian process regression can be described as

$$\log p(Y|X) = -\frac{1}{2}Y^\top(\mathbf{K}_{f,f} + \sigma^2 I)^{-1}Y - \frac{1}{2}\log|\mathbf{K}_{f,f} + \sigma^2 I| - \frac{nD}{2}\log 2\pi, \tag{2.115}$$

where D is the training data set dimension. Finally, we can describe the joint probability distribution in the form

$$p(Y, \mathbf{f}) = p(Y|\mathbf{f})p(\mathbf{f}). \tag{2.116}$$

This expression represents a non-parametric model as the dimension of the covariance matrix $\mathbf{K}_{f,f}$ increases with the size of the training set. In this review, I shall examine the Gaussian process regression model for predicting an output function $\mathbf{f}_*$ given a test input $\mathbf{x}_*$.

2.10.1.1 Covariance functions

In general, the relationship among observations is related one to another through the covariance matrix $\mathbf{K}_{f,f}$. A common choice of the covariance function is the *squared exponential* (SE) covariance function, sometimes known as the *radial basis function* (RBF) function, such that

$$\mathrm{Cov}(f(x_i), f(x_j)) = k(x_i, x_j) = \sigma_{rbf}^2 \exp\left\{-\frac{(x_i - x_j)^2}{2l^2}\right\} + \sigma^2\delta_{ij}, \tag{2.117}$$

where σ_{rbf}^2 is the signal variance that can amplify or reduce the signal for large or small values respectively. However, l is the characteristic length-scale that causes the function to vary rapidly or slowly when adjusting it for small or large values respectively, and $\epsilon \sim \mathcal{N}(0, \sigma^2\delta_{ij})$ is a Gaussian white noise. The covariance function determines the correlation between two random variables. For instance, the correlation between $f(x_1)$ and $f(x_2)$ is higher than the correlation between $f(x_1)$ and $f(x_5)$. In other words, if x_5 is farther from x_1 and x_2 the correlation

tends to zero as the separation distance increases. Moreover, this equation shows that the covariance of outputs is written in function of inputs. The choice of the covariance function $k(.,.)$ has to satisfy *Mercer's* theorem which states that any continuous, symmetric, positive semi-definite kernel function $\mathbf{K}$ can be expressed as a dot product in a high-dimensional space. It turns out that the covariance function is equivalent to the kernel function, as it is known in the *support vector machine*. By satisfying this condition one may develop many other forms of covariance functions such as Matérn, rational quadratic and so forth. Additionally, Rasmussen and Williams (2006) showed that one can create new valid kernels from old ones by using the sum, product and convolution operations.

2.10.1.2 Parameter estimation

In the machine learning literature, the parameters of a GP are usually estimated by maximising the marginal likelihood, which is also known as the type II maximum likelihood.
We re-write the marginal likelihood expression described in Equation (2.115) by conditioning it explicitly on the parameters of the covariance function, such that

$$\log p(Y|X,\boldsymbol{\theta}) = -\frac{1}{2}Y^T(\mathbf{K}_{f,f}+\sigma^2 I)^{-1}Y - \frac{1}{2}\log|\mathbf{K}_{f,f}+\sigma^2 I| - \frac{nD}{2}\log 2\pi. \quad (2.118)$$

The parameters are normally maximised with respect to each element on $\boldsymbol{\theta}$ using a gradient-descent method, which gives point estimates of the parameter vector $\boldsymbol{\theta}$.

2.10.1.3 Regression

The Gaussian process in regression would initially require the definition of the covariance function in a general way. For instance, let us assume that our physical process can best be described by an underlying RBF kernel, which is a smooth kernel function because the exponential function can have an infinitely large Taylor series expansions. We then inject a few observations or training points $\{X, \mathbf{f}(X)\}$ to define the covariance matrix among all possible combinations of these points,

so that

$$\mathbf{K}_{\mathbf{f},\mathbf{f}} = \begin{pmatrix} k(x_1,x_1) & k(x_1,x_2) & \cdots & k(x_1,x_n) \\ k(x_2,x_1) & k(x_2,x_2) & \cdots & k(x_2,x_n) \\ \vdots & \vdots & \ddots & \vdots \\ k(x_n,x_1) & k(x_n,x_2) & \cdots & k(x_n,x_n) \end{pmatrix}. \tag{2.119}$$

Combining the covariance function with the training set would result in defining the posterior mean, which should be used for prediction. If there are n_* test points, the prediction would consist of conditioning the joint Gaussian prior distribution on the observations, $p(\mathbf{f}_*|\mathbf{f})$, which would require us to evaluate the following covariance matrices

$$\mathbf{K}_{*,\mathbf{f}} = \begin{pmatrix} k(x_*^1,x_1) & \cdots & k(x_*^1,x_n) \\ k(x_*^2,x_1) & \cdots & k(x_*^2,x_n) \\ \vdots & \ddots & \vdots \\ k(x_*^{n_*},x_1) & \cdots & k(x_*^{n_*},x_n) \end{pmatrix} \text{ and } \mathbf{K}_{*,*} = \begin{pmatrix} k(x_*^1,x_*^1) & \cdots & k(x_*^1,x_*^{n_*}) \\ k(x_*^2,x_*^1) & \cdots & k(x_*^2,x_*^{n_*}) \\ \vdots & \ddots & \vdots \\ k(x_*^{n_*},x_*^1) & \cdots & k(x_*^{n_*},x_*^{n_*}) \end{pmatrix}, \tag{2.120}$$

where $\mathbf{K}_{*,\mathbf{f}}$ is the cross-covariance between $\mathbf{f}_*$ and $\mathbf{f}$, and $\mathbf{K}_{*,*}$ is the covariance function associated with $\mathbf{f}_*$. Accordingly, the complete form of the covariance matrix can be written in the form

$$\mathbf{K} = \begin{pmatrix} \mathbf{K}_{\mathbf{f},\mathbf{f}} & \mathbf{K}_{\mathbf{f},*} \\ \mathbf{K}_{*,\mathbf{f}} & \mathbf{K}_{*,*} \end{pmatrix}. \tag{2.121}$$

By multiplying the Gaussian likelihood, Equation (2.113), by the Gaussian process prior, Equation (2.112), we obtain the posterior distribution evaluated over $f(x)$, such that

$$\begin{aligned} p(f(x)|X,Y) \propto \mathcal{GP}(m_{post}(x) &= k(x,X)[\mathbf{K}(X,X)+\sigma^2 I]^{-1}Y, \\ k_{post}(x,x') &= k(x,x') - k(x,X)[\mathbf{K}(X,X)+\sigma^2 I]^{-1}k(X,x')). \end{aligned} \tag{2.122}$$

which represents a distribution over functions. To predict the function values $\mathbf{f}_*$ evaluated at different locations X_*, we would multiply the likelihood by the posterior and marginalise out the rest of the infinitely many variables $\mathbf{f}$ to keep the ones that we are interested in, such that

$$p(\mathbf{f}_*|X_*,X,Y) = \int \underbrace{p(\mathbf{f}_*|X_*,\mathbf{f})}_{Likelihood}\underbrace{p(\mathbf{f}|X,Y)}_{Posterior}\, d\mathbf{f}, \tag{2.123}$$

where the parameters of the model are the function $\mathbf{f}$ itself. For most interesting models this integral is intractable, but it becomes tractable in the GP framework. This is because the likelihood function is a Gaussian distribution at the observed points $f(x_1), f(x_2), \ldots, f(x_n)$ and the prior is a Gaussian process. The outcome would then be an infinite dimensional posterior Gaussian distribution, which is a distribution over functions. It turns out that Equation (2.123) becomes tractable because the integration of the product of two Gaussians is a Gaussian. Therefore, the predictive conditional distribution can be represented in the form

$$p(\mathbf{f}_*|X_*, Y, X) = \mathcal{N}(\bar{\mathbf{f}_*}, \mathbf{\Sigma}), \tag{2.124}$$

where

$$\bar{\mathbf{f}_*} = \mathbf{K}_{*,\mathbf{f}}\mathbf{K}_{\mathbf{f},\mathbf{f}}^{-1}\mathbf{f} \tag{2.125}$$

$$\mathbf{\Sigma} = \mathbf{K}_{*,*} - \mathbf{K}_{*,\mathbf{f}}\mathbf{K}_{\mathbf{f},\mathbf{f}}^{-1}\mathbf{K}_{\mathbf{f},*}. \tag{2.126}$$

$\bar{\mathbf{f}_*}$ is a linear representation of the predictive mean evaluated at the test points such that $\bar{\mathbf{f}_*} = \sum_{i=1}^{n} c_i k(x_*, x_i)$ and $\mathbf{\Sigma}$ defines the error bars on this prediction which is composed of two terms, $\mathbf{K}_{*,*}$ is the prior variance and $\mathbf{K}_{*,\mathbf{f}}\mathbf{K}_{\mathbf{f},\mathbf{f}}^{-1}\mathbf{K}_{\mathbf{f},*}$ is the covariance between the test points and the training set. The Equation (2.126) can be interpreted in the following way: the closer our test point is situated from the training point, the greater the $\mathbf{K}_{*,\mathbf{f}}\mathbf{K}_{\mathbf{f},\mathbf{f}}^{-1}\mathbf{K}_{\mathbf{f},*}$ expression would become, and hence the smaller the variance $\mathbf{\Sigma}$. However, the further our test point is situated from the training point, the smaller the $\mathbf{K}_{*,\mathbf{f}}\mathbf{K}_{\mathbf{f},\mathbf{f}}^{-1}\mathbf{K}_{\mathbf{f},*}$ expression would become, and therefore the larger the variance $\mathbf{\Sigma}$ we would obtain.
As we can see, the prediction performance is controlled entirely by the covariance matrix, which is an important key for regression.

2.10.2 Time series

A time series is a set of observations collected when observing the evolution of a physical process in time (e.g. finance, environmental, medicine). An autoregressive process (AR) of order p denoted as AR(p) can be defined in the form

$$X_t = \sum_{i=1}^{p} \phi_i X_{t-i} + \epsilon_t, \tag{2.127}$$

where $\{\phi_i\}_{i=1}^{p}$ are constant coefficients and ϵ_t is a sequence of uncorrelated white noise defined with mean 0 and variance σ^2. For a special case where $p = 1$, the AR(1) process takes the form

$$X_t = \phi_1 X_{t-1} + \epsilon_t. \tag{2.128}$$

Assuming stationarity, Weber (2007) described how one may determine the optimal parameters of an AR process. This starts by initially calculating the autocovariance function that is achieved by multiplying both sides of Equation (2.128) by X_{t-k} and then taking the expected value, such that

$$\begin{aligned} X_{t-k}X_t &= \phi_1 X_{t-k}X_{t-1} + X_{t-k}\epsilon_t \\ \Rightarrow \mathbb{E}(X_{t-k}X_t) &= \mathbb{E}(\phi_1 X_{t-k}X_{t-1}) + \mathbb{E}(X_{t-k}\epsilon_t) \\ \Rightarrow \mathbb{E}(X_{t-k}X_t) &= \phi_1 \mathbb{E}(X_{t-k}X_{t-1}) \\ \Rightarrow \gamma_k &= \phi_1 \gamma_{k-1}. \end{aligned}$$

The sequence γ_k (for $k = 1, 2, \ldots$) is called the autocovariance function. The autocorrelation function is defined as follows:

$$\rho_k = \gamma_k/\gamma_0 = \text{corr}(X_{t-k}X_t).$$

Similarly, squaring Equation (2.128) and taking the expected value as well as assuming covariance stationarity meaning that $\mathbb{E}(X_{t-1}^2) = \mathbb{E}(X_t^2)$, the variance can be directly computed,

$$\begin{aligned} \mathbb{E}(X_t^2) &= \phi_1^2 \mathbb{E}(X_{t-1}^2) + 2\phi_1 \mathbb{E}(X_{t-1}\epsilon_t) + \mathbb{E}(\epsilon_t^2) \\ \mathbb{E}(X_t^2) &= \phi_1^2 \mathbb{E}(X_{t-1}^2) + 0 + \sigma^2 \\ \mathbb{E}(X_t^2) &= \phi_1^2 \mathbb{E}(X_t^2) + \sigma^2 \\ \Rightarrow \gamma_0 &= \phi_1^2 \gamma_0 + \sigma^2 \\ \Rightarrow \gamma_0 &= \sigma^2/(1 - \phi_1^2). \end{aligned}$$

The autocorrelation function is obtained by multiplying Equation (2.127) by X_{t-k} and dividing it by γ_0, such that

$$\rho_k = \phi_1 \rho_{k-1} + \phi_2 \rho_{k-2} + \ldots + \phi_p \rho_{k-p} + \epsilon_t, \tag{2.129}$$

which is known as the Yule-Walker equation. The autocorrelation should be zero for distant observations, $\gamma_k \to 0$ as $k \to \infty$, in which case the optimal parameters $\phi_1, \phi_2, \ldots, \phi_p$ can be solved using a Levinson-Durbin recursion.
Moreover, the moving average process (MA) of order q denoted as MA(q), takes the form

$$X_t = \theta_0 + \sum_{i=1}^{q} \theta_i \epsilon_{t-i} + \epsilon_t, \tag{2.130}$$

where θ_0 and $\{\theta_i\}_{i=1}^{q}$ are the coefficients and ϵ_t is a sequence of uncorrelated white noise. A first order moving average process MA(1) has dynamics which follow

$$X_t = \theta_0 + \theta_1 \epsilon_{t-1} + \epsilon_t. \tag{2.131}$$

The mean and the variance are given by

$$\mathbb{E}(X_t) = \theta_0,$$
$$\mathbb{E}(X_t^2) = \sigma^2(1 + \theta_1^2).$$

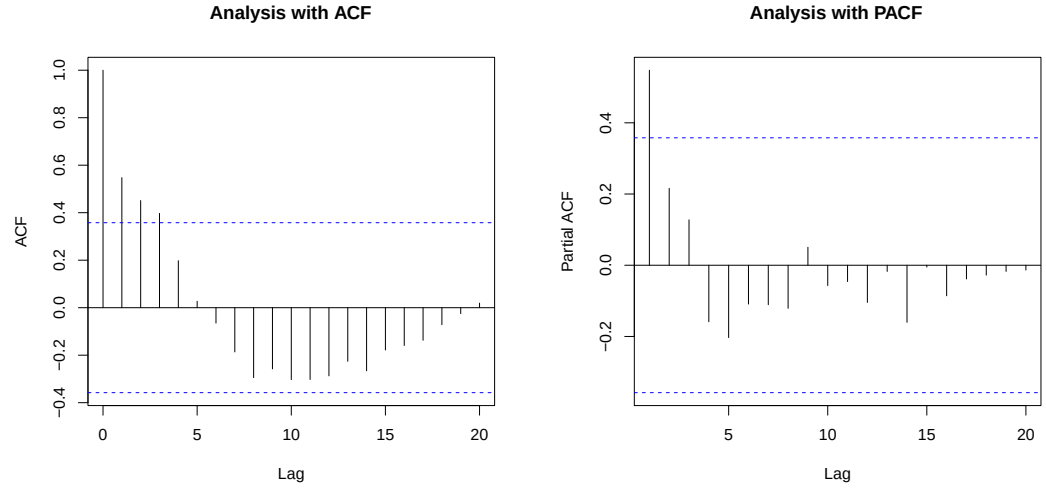
2.10.2.1 Autocorrelation function

The autocorrelation function (ACF) is commonly used to detect non-random patterns presented in a time series data. This technique consists of computing the autocorrelation at different time lags and analysing their values. Non-random patterns are generally identified by non-zero autocorrelations, in contrast to randomness which is identified by values near to zero. Figure 2.12-a illustrates the result of an ACF function applied to time series data, where the horizontal blue lines represent a 95% confidence interval and the bars represent the values. This figure tells us that the data is correlated such that current observation are depending on previous ones.

2.10.2.2 Partial autocorrelation function

The Partial autocorrelation function (PACF) is commonly used to define the order of an autoregressive model. It describes the autocorrelation function between X_t and X_{t-k} and disregards any information situated in between (e.g. lags 1 through $k-1$). For example, Figure 2.12-b illustrates the result of a PACF function applied to a time series data which shows a high autocorrelation value at lag 1. This recommends that an AR(1) would be a good option to describe the data.

Figure 2.12: Examples of Autocorrelation and Partial Autocorrelation functions.



(a) Autocorrelation function

(b) Partial Autocorrelation function

2.10.2.3 Autoregressive moving average

An autoregressive moving average model denoted as ARMA(p,q) is obtained by combining AR and MA processes, such that

$$X_t = \sum_{i=1}^{p} \phi_i X_{t-i} + \epsilon_t + \sum_{i=1}^{q} \theta_i \epsilon_{t-i}, \tag{2.132}$$

which is stationary for appropriate ϕ, θ. The contribution of Box and Jenkins (1976) was in developing a systematic procedure to determine the most appropriate values of p and q, which consists of three stages

1. *Model identification*: consists of assessing whether the data is derived from a stationary process or not. For instance, we can difference the data to remove non-stationarity patterns or, for more convenience, transform it (e.g. using the log function). Then we choose p and q such that the ACF for an MA(q) is zero beyond lag q, and the PACF for an AR(p) is zero beyond lag p.

2. *Parameter estimation*: consists of determining model parameters ϕs and θs using Levinson-Durbin recursion which restricts results to be within the unit circle. Accordingly, the approximate log likelihood can be defined by

$$-2 \log L = \sum_{t=1}^{n} \left\{ \log(2\pi) + 2 \log \sigma_{t-1} + \frac{(X_t - \mu_t)^2}{\sigma_{t-1}^2} \right\}, \tag{2.133}$$

where μ_t and σ_t^2 are functions of the parameters ϕs and θs. Thus, p and q can be chosen by minimising the AIC, such that

$$\text{AIC} = -2\log L + 2(p+q),$$

where $p+q$ is the number of unknown parameters in the model.

3. *Diagnostic checking*: consists of applying overfitting and residual analysis tests. The former consists of increasing the number of parameters gradually until we detect that further parameters are not reducing the AIC value; however, the latter consists of calculating the residuals from the fitted model and verifying that they are consistent with white noise. Ljung and Box (1978) applied a statistical test to check whether the overall randomness based on a number of lags is different from zero, known as the 'portmanteau' test of white noise. It is commonly applied to the residuals of a fitted ARMA model to test the hypothesis that residuals are uncorrelated.

Additionally, an autoregressive integrated moving average model denoted as ARIMA(p,d,q) is a generalisation of an ARMA model but with a differencing term to discard non-stationarity from the data.

2.10.2.4 Autoregressive conditional heteroskedasticity

Autoregressive conditional heteroskedasticity (ARCH) models (Engle, 1982) are used when we believe that the variance of a time series is not constant at every point in a series. Such models are often used in financial time series in which they assume that the current variance is expressed as a function of previous innovations. Accordingly, the variance of an ARCH(q) model takes the form

$$\sigma_t^2 = \alpha_0 + \alpha_1\epsilon_{t-1}^2 + \cdots + \alpha_q\epsilon_{t-q}^2 = \alpha_0 + \sum_{i=1}^{q}\alpha_i\epsilon_{t-i}^2, \tag{2.134}$$

where $\alpha_i \geq 0$. On the other hand, Bollerslev (1986) proposed a generalized autoregressive conditional heteroskedasticity (GARCH) model which is a natural generalization of the ARCH model allowing for a much more flexible lag structure,

that is, defined in the form

$$\sigma_t^2 = \alpha_0 + \alpha_1 \epsilon_{t-1}^2 + \cdots + \alpha_q \epsilon_{t-q}^2 + \beta_1 \sigma_{t-1}^2 + \cdots + \beta_p \sigma_{t-p}^2 = \alpha_0 + \sum_{i=1}^{q} \alpha_i \epsilon_{t-i}^2 + \sum_{i=1}^{p} \beta_i \sigma_{t-i}^2. \tag{2.135}$$

It should also be noted that the generalisation from ARCH to GARCH models is similar to the generalisation from an autoregressive (AR) model to autoregressive moving average (ARMA) model.

2.11 Summary

In this chapter, I reviewed some technical ground about probability theory: frequentist and Bayesian inferences. Bayesian inference theories have been widely used in artificial intelligence and expert systems since the late 1950s; but, unfortunately, the solution obtained by these models is rarely achievable in an analytical closed form. To simplify the computation of the posterior distribution, one may recourse to exponential families on the choice of priors; but for the majority of cases the posterior distributions are approximated by functional and MCMC approximation techniques. Since graphical models are at the heart of every probability model, I reviewed some basic concepts of graphical models that are necessary in understanding the probabilistic models and inference algorithms. Finally I discussed the advantage of Gaussian Process models and how one may fit a time series model to a process that evolves over time.

Chapter 3

Identification and quantification of heteroscedasticity in stock-recruitment relationships

Non-constant variance (heteroscedasticity) in the stock-recruitment (S-R) relationship is proposed as an important factor in sustainable fisheries management, but its reliable estimation from noisy populations is problematic. I developed methods for both frequentist and Bayesian approaches to test whether I can accurately estimate the degree of heteroscedasticity in 90 published S-R populations. The confidence interval for the heteroscedastic regression model is estimated via a parametric bootstrap approach, and the credible interval for the Bayesian method via a Markov chain Monte Carlo sampling algorithm. I found strong evidence of negative heteroscedasticity in several stocks, regardless of the statistical paradigm, the details of density dependence, and the methods used to generate the original populations. This statistical framework provides an efficient and reliable setting for assessing heteroscedasticity of the S-R relationship in fisheries.

The objectives of this Chapter are the following:

- To examine whether the additional parameter η_1 can provide a better fit to the data.
- To examine whether the parameter η_1 can be reliably estimated.

Much of the work in this chapter has recently been published by the author and colleagues (Panikian et al., 2015), printed in Appendix H.

3.1 Introduction

Reliable mathematical modelling and prediction of fish populations is of great importance socially and economically, as well as being a necessary ingredient in the conservation of biodiversity. Various natural and anthropogenic factors affect fish populations, with the life of juvenile fish typically being characterised by enormous mortality rates (Hilborn and Walters, 1992). Newly hatched fish larvae have very low probability of reaching adulthood (Pitchford et al., 2005). Mortality is due to variability in food supply, migration, predation, starvation, poisonous pollutants, and fishing activities (Steele et al., 1977): resulting in an unpredictable relationship between the adult population ('stock') and the juveniles ('recruitment') that will successfully survive to enter the adult population in the future. Understanding the stock-recruitment (S-R) relationship therefore requires careful statistical techniques forming a crucial ingredient in the sustainable management of these exploited natural resources.

There are, of course, limits to growth in populations. For instance, climatic changes, environmental conditions and natural disasters are classified as density-independent factors that influence larval survival and recruitment size directly but do not regulate variability of juvenile mortality, except for some flatfish species (Myers and Cadigan, 1993b; Leggett and Deblois, 1994). In contrast, intraspecific competition, predation and disease are classified as density-dependent limiting factors that might regulate the recruitment variation of fish populations (Myers and Cadigan, 1993a). At low population there would be very little density-dependent mortality during the juvenile stage, but when the population size increases a strong density-dependent mortality usually occurs. Myers and Cadigan (1993b) found that the interannual variability in juvenile survival appears to be the most important source of variability in abundance; but it is attenuated by density-dependent mortality in the juvenile stage. From another perspective, Spencer (2008) studied the effect of both density-dependent and density-independent factors in determining the spatial distribution of six flatfish species living in the eastern Bering Sea. This distribution is found to be shifting northward toward colder habitats in response to increasing temperatures caused by global warming.

Understanding the stock-recruitment (S-R) relationship therefore requires careful statistical techniques forming a crucial ingredient in the sustainable management

of these exploited natural resources. From the point of view of sustainable management, Shepherd and Cushing (1990) studied plausible regulatory processes for analysing fish populations and argued that increased variability at low stock sizes might prevent the collapse of stocks subject to high mortality rates, because in this case the variability acts to produce depensatory rather than compensatory density-dependence, a theme echoed by Minto et al. (2008). Hsieh et al. (2006) presented the first empirical evidence that fishing could increase the survival variability (a proxy for recruitment variability) in an exploited population and advocated that increased variability of exploited populations favours a precautionary management approach.

Heteroscedastic models (i.e. statistical models using non-constant variance) have gained much interest in recent years to explain the regulatory mechanisms in fish populations. Minto et al. (2008) developed a stochastic method applied to a meta-analysis of 147 fisheries populations to argue that survival variability is inversely proportional to stock size. Their model was inspired by Peterman (1981) who argued that random variation in marine survival rates tends to follow a log-normal distribution; but the novelty of their method was to incorporate a functional form of non-constant recruitment variability over adult abundance. More recently, Burrow et al. (2012) investigated the feasibility of applying heteroscedastic models in practice, using two North Sea stocks as examples. They uncovered a weakness of using a heteroscedastic regression model by showing it to be statistically unreliable to fit the parameters based on small S-R populations (containing 40 or 50 data points); but made a mistake while defining the log-likelihood function (i.e. missing a square term and a factor of 0.5) and restricting their analysis to only two populations. The use of heteroscedastic models is controversial because previous research engaged in interpreting non-constant variance has failed to provide a clear-cut answer about its reliable estimation for fisheries management.

The aims of this study were: (1) to develop frequentist and Bayesian methods for accurately identifying the non-constant variance exhibited in a density-dependent model, and (2) to test the reliability of these methods on 90 S-R populations. Since none of the S-R populations are direct observations, I select populations estimated by virtual population analysis (VPA) type assessments so as to ensure that the recruitment estimates are derived from the catch-at-age data, which is not dependent on the estimate of the spawning stock biomass. I found it useful

to analyse the edge effects at the beginning and end of the time series data to test whether VPA methods have an impact upon our results. In this work, I employed the two dominant approaches to inference, known as Bayesian and frequentist statistical methods, to determine whether one can reliably estimate the non-constant variance. I conclude that within either the frequentist or Bayesian paradigm, the reliability of determining the existence of a negative η_1 values (the coefficient of heteroscedasticity) can only be assessed on a case-by-case basis.

3.2 Materials and methods

I pruned S-R populations collated in the publicly available RAM legacy database (www.ramlegacy.org) (Ricard et al., 2012) by restricting the analysis only to those estimated by virtual population analysis (VPA) type assessments. The spawning stock biomass (SSB) is measured in tonnes; however, the recruitment is measured in thousands of individuals. The 12 VPA-type assessment methods classified under this category are as follows: VPA, SPA, XSA, FLXSA, ADAPT, NFT-ADAPT, B-ADAPT, SXSA, SPA-ADAPT, NFT-ADAP, ISVPA and hybrid. VPA, also known as cohort analysis, follows cohorts through their whole life, using catch-at-age data and natural mortality to back-calculate what recruitment had to be in order to support the catch (Hilborn and Walters, 1992). In contrast, assessments based on integrated analyses and statistical catch-at-age assessments employ an underlying S-R relationship, so fitting a S-R curve to their time-series is not appropriate. There were 100 S-R populations obtained with VPA-type assessment, but 10 populations had missing data or no data at all. Accordingly, I restricted the analysis on the remaining 90 fish populations, representing 32 species (see Appendix A, Table A.1).

For the purposes of illustration, here I briefly describe a simple VPA analysis from (Anderson, 1978) to show how one can estimate stock sizes and fishing mortality rates for each year-class (cohort) making up the overall population. The approach relies on two simple relationships for each cohort, such that

$$C_i = N_i \frac{F_i}{M_i + F_i} \left\{1 - \exp(-M_i - F_i)\right\}, \tag{3.1}$$

$$\frac{N_{i+1}}{C_i} = \frac{(M_i + F_i)\exp(-M_i - F_i)}{F_i \left\{1 - \exp(-M_i - F_i)\right\}}, \tag{3.2}$$

where C_i is the catch of fish of that cohort in year i, N_i is the population size at the beginning of year i, M_i is the natural mortality rate (estimated independently from previous research) and F_i is the fishing mortality. Assume (for example) that no individual exceeds an age of nine years (or equivalently, one can introduce a plus group for individuals aged 9 and above). Then by knowing C_i, M_i and by estimating F_i at age nine by incorporating ancillary data in the form of tagging experiments (Parks, 1976), we would be able to iteratively calculate the population size each year, starting from the oldest and moving backward to the youngest. Explicitly N_9 is solved using Equation (3.1). Next, the fishing mortality at age eight or F_8 is solved using Equation (3.2). This result is substituted into the catch equation (3.1) to calculate N_8, and so forth down to the age one. The outcome of the VPA analysis is then used to estimate recruitment expressed as abundance at age 1 and the SSB by summing up stock sizes of age 2+ in each year respectively. There are many variations on this basic theme, which use age-structured data to estimate current stock size. I confirmed that my statistical results are not affected by the VPA-type assessment, and hence I assume that all relevant S-R datasets are approximately derived from the same underlying model.

3.2.1 The Model

To understand the relationship between spawning stock biomass (SSB) and recruitment, Minto et al. (2008) proposed the following model:

$$\ln\left(\frac{R_i}{S_i}\right) \sim_{\text{i.i.d.}} \mathcal{N}(\mu_i, \sigma_i^2) \quad \text{where} \quad \mu_i = \ln(\alpha) + \frac{1}{\gamma}\ln(1-\gamma\beta S_i) \quad \text{and} \quad \sigma_i^2 = \exp(\eta_0 + \eta_1 S_i), \tag{3.3}$$

where R_i and S_i are the estimated number of recruits and SSB for each observation i respectively. The parameter γ is fixed during the analysis; but α, β, η_0 and η_1 need to be estimated. This is a regression model that assumes the logarithm of the ratio (R_i/S_i) is an independent and identically distributed (i.i.d.) sample from a Gaussian model with non-constant variance. In practice, none of the populations (i.e. spawning stock biomass and recruitment) are actually direct observations. They are in reality model outputs (parameter estimates) from fisheries assessments, where models have been previously fitted to fisheries data

(catch, age structure information, indices, etc.). The parameters α and β measure the productivity and the density-dependent mortality (capacity) in a population, respectively. The density-independent part of the variance is described by η_0, with the density dependent variance described by η_1, known as the heteroscedastic coefficient. The parameter γ enables us to choose between several survival models. For instance, $\gamma = -1000$ generates a model with essentially no density dependence, and increasing γ increases the amount of density dependence: reproducing several models that have been advocated in previous studies (Minto et al., 2008), such as: $\gamma = -2$ (Cushing-like), $\gamma = -1$ (Beverton-Holt), $\gamma = 0$ (Ricker) and $\gamma = 1$ (Schaefer).

3.2.2 Likelihood Of The Model

Let $\boldsymbol{R} = (R_1, R_2, \ldots, R_n)$ and $\boldsymbol{S} = (S_1, S_2, \ldots, S_n)$ be the recruitment and stock model input vectors. The log-likelihood of the heteroscedastic regression model is

$$\mathcal{L}\{\ln(\boldsymbol{R}/\boldsymbol{S}), \alpha, \beta, \eta_0, \eta_1\} = -\frac{n}{2}\ln(2\pi) - \frac{1}{2}\sum_{i=1}^{n}(\eta_0 + \eta_1 S_i) - \frac{1}{2}\sum_{i=1}^{n}\frac{\left\{\ln\left(\frac{R_i}{S_i}\right) - \mu_i\right\}^2}{\exp(\eta_0 + \eta_1 S_i)}, \tag{3.4}$$

where

$$\mu_i = \ln(\alpha) + \frac{1}{\gamma}\ln(1 - \gamma\beta S_i),$$

and n is the number of observations. For the initial set of experiments, I fix $\gamma = -1$ for my analyses here (representing the Beverton-Holt compensation model (Beverton and Holt, 1957)), which turns the log-likelihood function into the form

$$\mathcal{L}\{\ln(\boldsymbol{R}/\boldsymbol{S}), \alpha, \beta, \eta_0, \eta_1\} \propto -\frac{1}{2}\sum_{i=1}^{n}(\eta_0 + \eta_1 S_i) - \frac{1}{2}\sum_{i=1}^{n}\frac{\left\{\ln\left(\frac{R_i}{S_i}\right) - \ln(\alpha) + \ln(1 + \beta S_i)\right\}^2}{\exp(\eta_0 + \eta_1 S_i)}. \tag{3.5}$$

For a constant variance, the coefficient of heteroscedasticity is zero and the variance would be written as $\sigma^2 = e^{\eta_0}$. I choose to scale both SSB and recruitment model inputs with their maximum values respectively so as to normalise the assessments between 0 and 1, as in (Minto et al., 2008).

To determine whether the log-likelihood function for Equation (3.5) is globally concave or not, I examined the matrix of second derivatives (or the Hessian

matrix):

$$\mathrm{H}(\alpha,\beta,\eta_0,\eta_1)=\begin{bmatrix} \dfrac{\partial^2\mathcal{L}}{\partial\alpha^2} & \dfrac{\partial^2 L}{\partial\beta\partial\alpha} & \dfrac{\partial^2\mathcal{L}}{\partial\eta_0\partial\alpha} & \dfrac{\partial^2\mathcal{L}}{\partial\eta_1\partial\alpha} \\ \dfrac{\partial^2\mathcal{L}}{\partial\alpha\partial\beta} & \dfrac{\partial^2\mathcal{L}}{\partial\beta^2} & \dfrac{\partial^2\mathcal{L}}{\partial\eta_0\partial\beta} & \dfrac{\partial^2\mathcal{L}}{\partial\eta_1\partial\beta} \\ \dfrac{\partial^2\mathcal{L}}{\partial\alpha\partial\eta_0} & \dfrac{\partial^2\mathcal{L}}{\partial\beta\partial\eta_0} & \dfrac{\partial^2\mathcal{L}}{\partial\eta_0^2} & \dfrac{\partial^2\mathcal{L}}{\partial\eta_1\partial\eta_0} \\ \dfrac{\partial^2\mathcal{L}}{\partial\alpha\partial\eta_1} & \dfrac{\partial^2\mathcal{L}}{\partial\beta\partial\eta_1} & \dfrac{\partial^2\mathcal{L}}{\partial\eta_0\partial\eta_1} & \dfrac{\partial^2\mathcal{L}}{\partial\eta_1^2} \end{bmatrix}. \tag{3.6}$$

To simplify the expressions I set $\xi_i=\ln(\frac{R_i}{S_i})-\ln(\alpha)+\ln(1+\beta S_i)$, $a_i=\exp(\eta_0+\eta_1 S_i)$ and $b_i=\frac{S_i}{1+\beta S_i}$. Thus, the Hessian matrix would take the form

$$\mathrm{H}(\alpha,\beta,\eta_0,\eta_1)=\begin{bmatrix} -\frac{1}{\alpha^2}\sum_{i=1}^n\frac{(1+\xi_i)}{a_i} & \frac{1}{\alpha}\sum_{i=1}^n\frac{b_i}{a_i} & -\frac{1}{\alpha}\sum_{i=1}^n\frac{\xi_i}{a_i} & -\frac{1}{\alpha}\sum_{i=1}^n\frac{S_i\xi_i}{a_i} \\ \frac{1}{\alpha}\sum_{i=1}^n\frac{b_i}{a_i} & \sum_{i=1}^n b_i^2\frac{(\xi_i-1)}{a_i} & \sum_{i=1}^n b_i\frac{\xi_i}{a_i} & \sum_{i=1}^n b_i S_i\frac{\xi_i}{a_i} \\ -\frac{1}{\alpha}\sum_{i=1}^n\frac{\xi_i}{a_i} & \sum_{i=1}^n b_i\frac{\xi_i}{a_i} & -\frac{1}{2}\sum_{i=1}^n\frac{\xi_i^2}{a_i} & -\frac{1}{2}\sum_{i=1}^n S_i\frac{\xi_i^2}{a_i} \\ -\frac{1}{\alpha}\sum_{i=1}^n\frac{S_i\xi_i}{a_i} & \sum_{i=1}^n b_i S_i\frac{\xi_i}{a_i} & -\frac{1}{2}\sum_{i=1}^n S_i\frac{\xi_i^2}{a_i} & -\frac{1}{2}\sum_{i=1}^n S_i^2\frac{\xi_i^2}{a_i} \end{bmatrix}, \tag{3.7}$$

from which I derived the sequence of determinants known as *principal minors*.

Definition 3.2.1. Let A be an $n\times n$ matrix; a $k\times k$ submatrix of A formed by deleting $n-k$ rows of A, and the same $n-k$ columns of A, is called principal submatrix of A. The determinant of a principal submatrix of A is called a principal minor of A.

Because the principal minors do not have a simplified form, a numerical solution becomes essential to test the alternation in sign of principal minors. To do so, I defined a subset $C\subset\mathbb{R}^4$ such that $C=\{0<\alpha<10, 0<\beta<10, -5<\eta_0<10, -5<\eta_1<10\}$ from which I randomly sampled six sets of parameters $(\alpha,\beta,\eta_0,\eta_1)$ and plugged them in the log-likelihood function. Generally speaking, a function is said to be concave if it satisfies the following theorem:

Theorem 1. *A twice differentiable real-valued function defined on an open convex set C is **concave** if and only if the Hessian matrix is negative semi-definite everywhere on C. In other words, if and only if $(-1)^k\Delta_k\geq 0$ for $k=1,2,\ldots,n$. At any point, the leading principal minors must alternate in sign with $\Delta_1\leq 0$, $\Delta_2\geq 0$, $\Delta_3\leq 0$ and so forth (Boyd and Vandenberghe, 2004).*

Table 3.1: Random selection of parameters $(\alpha, \beta, \eta_0, \eta_1)$ with their associated principal minors (Δ_1, Δ_2, Δ_3, Δ_4) obtained while analysing the HAWG-HERRVIaVIIbc-1956-2010 stock-recruitment dataset. The first four columns describe the sampled parameters, but the last four columns describe the corresponding principal minors.

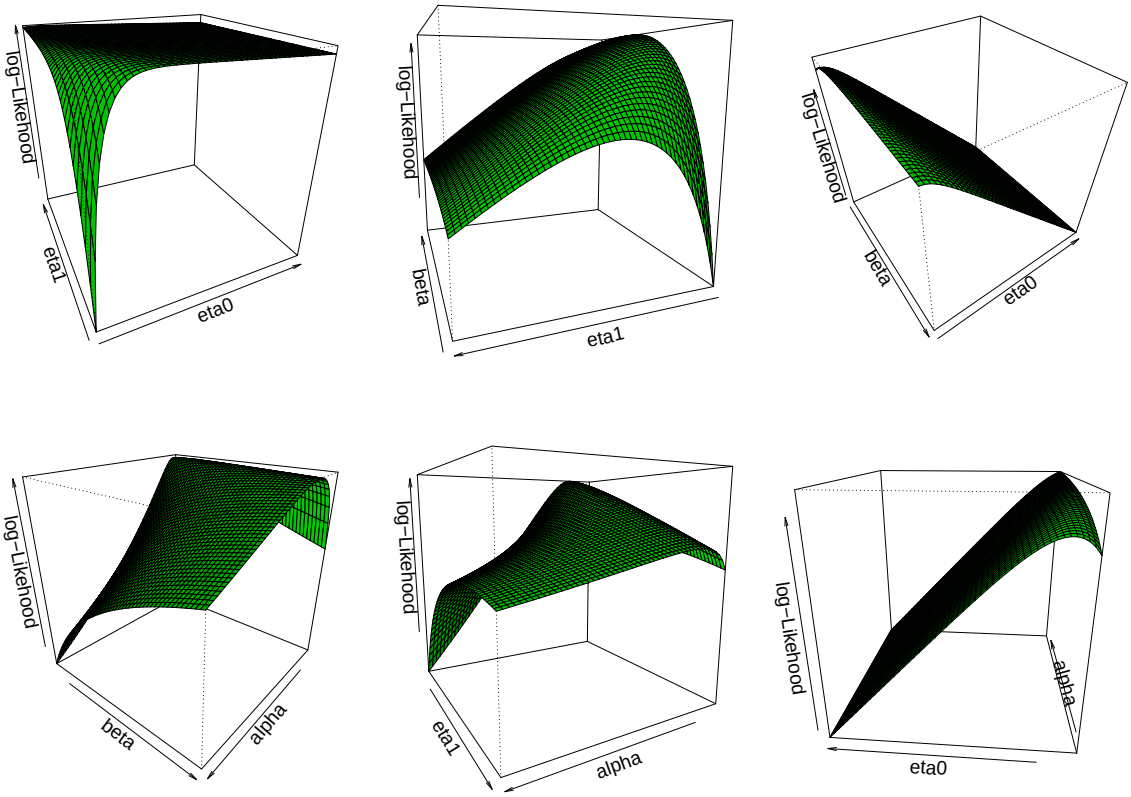
Sampled Parameters				Principal Minors			
α	β	η_0	η_1	Δ_1	Δ_2	Δ_3	Δ_4
4.658601	3.110526	-3.096537	6.527564	2.607741	-55.056882	16298.42608	-37216.55418
8.910902	1.921319	3.00684	3.99925	1.24E-02	-1.06E-03	2.71E-03	-9.76E-05
0.03391379	2.79001425	1.34757225	4.18833674	-19159.197	-4166.447	128401.969	-67830.031
3.3805582	8.0489839	0.8670313	5.646247	-2.63E-01	-2.73E-03	5.56E-03	-8.68E-05
8.452809	9.76857	8.68387	1.497699	2.12E-05	-3.55E-09	2.13E-11	-2.95E-15
2.886448	6.962125	9.985499	7.325011	-3.40E-05	-1.47E-11	2.37E-15	-1.96E-21

The *principal minors* are found to not alternate in sign (Table 3.1) as they showed: (1) *local concavity* (first and second rows), (2) *indefinite curvatures* described by the arbitrary change in the sign of principal minors (third and fourth rows) and (3) *flat curvatures* described by the null principal minors (last two rows). Since there is at least one sample that disagrees with Theorem 1, the log-likelihood function is not concave and requires suitable methods for solving this optimization problem. Figure 3.1 illustrates the shape of the log-likelihood function that is obtained by projecting it on all possible dimensional spaces.

3.2.3 Why choose a heteroscedastic regression model?

I compared model fitting for heteroscedastic and non-heteroscedastic regression models using the AICc statistic, I found a prevalence of the heteroscedastic model for 78 out of 90 populations showing that the heteroscedastic model had a much better fit across the majority of stocks, regardless of the coefficient of heteroscedasticity. Since the sign of η_1 has a great influence in determining whether such a model is appropriate for devising optimal harvest strategies, I pursue an inquiry in the remainder of this Chapter to investigate whether the sign of η_1 can be reliably estimated. I also applied the AICc statistic as a measure to determine the most appropriate value for γ that fits the S-R populations.

Figure 3.1: Projection of the log-likelihood function on all possible axes.



3.2.4 Frequentist inference

In this work, I apply maximum likelihood estimation (MLE) method to find a specific estimate $\hat{\boldsymbol{\theta}}$ of $\boldsymbol{\theta} = (\alpha, \beta, \eta_0, \eta_1)$ that maximises the log-likelihood function $\mathcal{L}(\boldsymbol{\theta}|\boldsymbol{R}, \boldsymbol{S})$. If there is a unique maximum, then the MLE estimator $\hat{\boldsymbol{\theta}}$ is consistent and asymptotically normal with its mean concentrated near the true value $\boldsymbol{\theta}$. Let $\hat{\boldsymbol{\theta}}$ be the maximum likelihood estimator (MLE) of $\boldsymbol{\theta}$ that maximises the log-likelihood function $\mathcal{L}(\boldsymbol{R}, \boldsymbol{S}|\boldsymbol{\theta})$. Here, I investigate several optimisation techniques. I first develop a Simulated Annealing (SA) algorithm with a uniform perturbation proposal distribution inspired from (Robert and Casella, 2009, p.144), the pseudocode example is described in Algorithm 5.

I additionally employ Nelder-Mead (Nelder and Mead, 1965), quasi-Newton (Byrd et al., 1995) and conjugate-gradient (Fletcher and Reeves, 1964) optimisation algorithms available in the built-in R function `optim`. Furthermore, I employ the AD Model Builder (ADMB) that is a free software package designed to help

Algorithm 5 The implemented SA algorithm, adapted from (Robert and Casella, 2009).

Define the domain of definition of parameters.
Define the number of random restarts and the number of iterations.
for $rand \leftarrow 1, n.Rand$ **do**
    Initialise the $Temp$, $Factor$ and $Scale$ parameters to 1.
    Initialise $\boldsymbol{\theta} = (\alpha, \beta, \eta_0, \eta_1)$ to $\boldsymbol{\theta}_0$.
    Store the initial step in $state.cur = \boldsymbol{\theta}_0$.
    Store the initial likelihood value in $curL = L(\boldsymbol{\theta}_0)$.
    for $iter \leftarrow 1, n.Iter$ **do**
        Propose a single step for each parameter.
        For example, $\alpha^* = state.cur[1] + runif(1, -1, 1) * Scale$.
        if $Temp * \log(runif(1)) < (L(\boldsymbol{\theta}^*) - curL)$ **then**
            $\boldsymbol{\theta} = \boldsymbol{\theta}^*$
            $curL = L(\boldsymbol{\theta}^*)$
        end if
        state.st[iter,] = state.cur #store the values.
        $Temp = 1/\log(1 + iter)$. # set the temperature schedule.
        acc.steps = length(unique(state.st[(ceiling(iter/2)):iter,1])).
        #in case of a continuous rejection, minimise the jumps.
        if $(acc.steps == 1)$ **then**
            $Factor = Factor/10$.
        end if
        #in case of a continuous acceptance, maximise the jumps to avoid local maxima.
        if $(2 * acc.steps > iter)$ **then**
            $Factor = Factor * 10$.
        end if
        Scale = $\max(2, Factor * \sqrt{Temp})$.
    end for
    For each random restart, store the parameters corresponding to the maximum likelihood in a global array.
end for
Find the largest maximum likelihood amongst the random restarts.

Table 3.2: Five optimisation techniques applied to two randomly selected datasets, AFWG-POLLNEAR-1957-2011 and NEFSC-HADGB-1930-2008 respectively. The maximised log-likelihood (max LL) value shows the strength of the algorithm and time(s) represents the elapsed time in seconds.

dataset	method	$\hat{\alpha}$	$\hat{\beta}$	$\hat{\eta}_0$	$\hat{\eta}_1$	max LL	time(s)
1	ADMB	9.417463	19.4767	-1.024932	-0.9773355	**14.86818**	2.09
	CG	5.707483	11.06662	-1.005155	-1.000677	14.67919	3
	Nelder-Mead	9.416474	19.47436	-1.024609	-0.9776983	**14.86818**	0.9
	quasi-Newton	9.302521	19.21712	-1.024472	-0.9779884	14.86811	1.15
	SA	9.990274	20.72245	-1.031662	-0.9748721	14.86555	461.94
2	ADMB	0.2826475	0.9606392	0.3339682	1.070824	**-62.16186**	4.66
	CG	0.2750472	0.8212238	0.3538997	0.9994656	-62.17271	3.22
	Nelder-Mead	0.2826419	0.9604135	0.3341409	1.070727	**-62.16186**	1
	quasi-Newton	0.2826537	0.9607349	0.33397	1.070809	**-62.16186**	0.51
	SA	0.285909	1.036674	0.3617885	1.010243	-62.16706	496.09

Table 3.3: Comparison between (a) ADMB and quasi-Newton, and (b) ADMB and Nelder-Mead, when applied onto 90 S-R datasets.

	Compared to	
	quasi-Newton	Nelder-Mead
ADMB =	38	33
ADMB >	45	40
ADMB <	7	17

ecologists solve numerous statistical problems (Fournier et al., 2012). All these methods are applied on two representative datasets so as to maximise the log-likelihood function, as described in Equation (3.5). Table 3.2 illustrates the point estimates of each optimisation technique along with the maximised log-likelihood and elapsed time values respectively. All these techniques provide very close estimate parameters such that the best method is identified as the one that gives the largest log-likelihood value. In these two cases, CG performs the worst amongst. Because of the Metropolis algorithm's, accept-reject proposals, SA is found to suffer from high computational cost which makes it impractical to be applied to all the datasets. To evaluate quasi-Newton, Nelder-Mead and ADMB, I applied these methods to all the datasets where I found the ADMB to be the most appropriate optimisation technique for this problem. Table 3.3 summarizes this comparison and shows that ADMB performs better than both quasi-Newton and Nelder-Mead techniques. For instance, I found that ADMB and quasi-Newton are equal over 38 cases, ADMB performs better than the quasi-Newton over 45 cases and worse over 7 cases.

Efron (1979) was the first to introduce the resampling residuals method to

construct a bootstrap empirical distribution from an original dataset assumed to be from an i.i.d. population. His method was applied to a series of examples including the basic case of a simple linear regression for which the method reliably assigned measures of accuracy of the model parameters. After his discovery many other researchers employed bootstrap methods in the literature as an approach for estimating the distribution of the estimator, see for example (Efron and Tibshirani, 1986). Efron's method assumes the distribution of the residual errors are approximately normally distributed. This normal approximation works well when the residual errors are homoscedastic, but can fail when residual errors are heteroscedastic making it inappropriate for estimating the parameters; see Shao (1988) and Wu (1986). A heteroscedastic bootstrap method was proposed by Wu (1986) for resampling residuals using a weighted bootstrap technique. Wu's method consists of slightly modifying Efron's original method to consider the nonconstancy of the error variances presented in the data. Liu (1988) proposed that Wu's sampling method could be substituted by randomly selecting samples from a population that has its third central moment equal to one with a zero mean and unit variance in order to obtain the second order properties of Wu's bootstrap. DiCiccio and Efron (1996) proposed a parametric bootstrap method with second order accuracy and correctness for producing good approximate confidence intervals, which has no distributional assumption on the mean-variance model.

To assess the estimation error for η_1 (heteroscedasticity parameter), I first employed the weighted residual bootstrap methods as used by Wu and Liu to investigate the properties of the estimator $\hat{\eta_1}$, but I found that these methods did not accurately approximate the 95% confidence interval for a particular dataset (namely INIDEP-ARGHAKENARG-1985-2007) because the MLE used for bootstrap simulations is found to be outside the approximate 95% confidence interval. As a result, I decided to abandon the weighted resampling residuals methods and use instead the parametric bootstrap sampling approach, as in DiCiccio and Efron (1996). This method is also known as bootstrapping raw data, where each replication is obtained by sampling from the heteroscedastic distribution fitted with the MLE $\hat{\boldsymbol{\theta}}$. The theory of this method shows that the bootstrap confidence intervals are second-order correct as well as second-order accurate (DiCiccio and Efron, 1996, sections 8 and 9) and it is appropriate for studies with small sample size. I describe this sampling method for simulating new recruits as follows:

Step 1: Use both stock and recruitment model inputs to estimate $\hat{\boldsymbol{\theta}} = (\hat{\alpha}, \hat{\beta}, \hat{\eta_0}, \hat{\eta_1})$.

Step 2: Draw i.i.d. samples $X^* = (x_1^*, x_2^*, \ldots, x_n^*)$ from $\mathcal{N}(\hat{\mu_i}, \hat{\sigma}_i^2)$ where

$$\hat{\mu_i} = \ln(\hat{\alpha}) - \ln(1 + \hat{\beta} S_i) \quad \text{and} \quad \hat{\sigma}_i^2 = \exp(\hat{\eta}_0 + \hat{\eta}_1 S_i),$$

with n is the number of data points found in the S-R population, and $i = 1, \ldots, n$.

Step 3: Simulate new recruit $\boldsymbol{R}^* = \boldsymbol{S} \exp(X^*)$, such that $\boldsymbol{R}^* = (R_1^*, R_2^*, \ldots, R_n^*)$.

Step 4: Scale $\boldsymbol{R}^*$ with its maximum value so as to range between 0 and 1.

Step 5: Re-fit the regression model to the simulated data $(\boldsymbol{R}^*, \boldsymbol{S})$ and estimate $\hat{\boldsymbol{\theta}}^*(\boldsymbol{R}^*|\boldsymbol{S}) = \arg\max_{\hat{\boldsymbol{\theta}}} \hat{\boldsymbol{\theta}}(\boldsymbol{R}^*|\boldsymbol{S})$.

Step 6: Repeat steps 2 to 5, 1000 times so as to obtain a good approximation of the confidence interval.

I compared the parametric bootstrap method with the asymptotic confidence intervals from ADMB for the subset of populations with adequate sample sizes (greater than 30) and positive definite Hessians; I found that the bootstrap method provides empirical coverages for $\hat{\eta}_1$ noticeably wider than the asymptotic confidence intervals (Table 3.4). The findings confirm that under the first order asymptotic theory the residual errors of recruitment are not normally distributed rendering the first order asymptotic theory inappropriate to assess the estimation error for η_1. To account for possible skewness of the estimator, Singh (1981) and DiCiccio and Efron (1996) proved that second-order properties are often more desirable as they improve by an order of magnitude upon the accuracy of the standard intervals.

During this analysis I found the Hessian matrix for three populations, namely INIDEP-PATGRENADIERSARG-1983-2006, NWWG-HADFAPL-1955-2011 and NWWG-HERRIsum-1984-2011 (also known by their id number: 25, 58 and 60), not positive definite meaning that the optimizer might fail to find the highest likelihood for which the parametric bootstrapping method would result in estimating incorrect MLEs. To overcome this hurdle, one can apply a Bayesian approach to estimate the posterior distribution of η_1, as discussed in Section 3.2.5.

The above analysis could be incomplete, because I focused only on the Beverton-Holt compensation model, see Equation (3.5). Here I generalise my previous assumption by making available the set of possible models $\gamma \in \{-2, -1, 0, +1\}$

Table 3.4: Descriptive comparison of asymptotic and bootstrap methods for estimating the approximate 95% confidence interval for η_1.

			Asymptotic 95% CI		Bootstrap 95% CI	
AssessId	noSamples	$\hat{\eta}_1$	Lower limit	Upper limit	Lower limit	Upper limit
AFWG-GHALNEAR-1960-2010	43	1.31	0.85	1.76	-1.24	3.16
AFWG-HADNEAR-1947-2010	58	-1.77	-2	-1.54	-3.42	0.31
AFWG-HADNS-IIIa-1963-2011	49	0.33	0.01	0.66	-2.4	2.72
AFWG-POLLNEAR-1957-2011	49	-0.98	-1.16	-0.79	-2.52	0.43
DFO-HAD5Zejm-1968-2003	34	1.56	1.12	1.99	-0.94	3.62
DFO-HERR4VWX-1964-2006	41	2.1	1.75	2.44	-0.27	4.05

and choose the one that provides the minimum AICc score, for each population respectively. A non-asymptotic recruitment is obtained for $\gamma = -2$, which means that recruitment can grow with adult abundance size. However, an overcompensation model is obtained for $\gamma \in \{-1, 0, +1\}$, which are different sorts of density dependence models. For $\gamma \in \{-2, -1, 0, +1\}$ I obtain different models but with the same number of parameters $\{\alpha, \beta, \eta_0, \eta_1\}$. As noted in Minto et al. (2008), I generated the log-likelihood functions of these models such as: the Cushing-like model is obtained for $\gamma = -2$,

$$\mathcal{L}\{\ln(\boldsymbol{R}/\boldsymbol{S}), \alpha, \beta, \eta_0, \eta_1\} \propto -\frac{1}{2}\sum_{i=1}^{n}(\eta_0+\eta_1 S_i) - \frac{1}{2}\sum_{i=1}^{n}\frac{\left\{\ln\left(\frac{R_i}{S_i}\right) - \ln(\alpha) + \frac{1}{2}\ln(1+2\beta S_i)\right\}^2}{\exp(\eta_0+\eta_1 S_i)}. \tag{3.8}$$

The Beverton-Holt model is obtained for $\gamma = -1$,

$$\mathcal{L}\{\ln(\boldsymbol{R}/\boldsymbol{S}), \alpha, \beta, \eta_0, \eta_1\} \propto -\frac{1}{2}\sum_{i=1}^{n}(\eta_0+\eta_1 S_i) - \frac{1}{2}\sum_{i=1}^{n}\frac{\left\{\ln\left(\frac{R_i}{S_i}\right) - \ln(\alpha) + \ln(1+\beta S_i)\right\}^2}{\exp(\eta_0+\eta_1 S_i)}. \tag{3.9}$$

The Ricker model is obtained after developing a first-order Taylor expansion of the $\log(1-\gamma\beta S_i)$ function around $\gamma \to 0$ resulting in $-\beta S_i$ after dividing it by γ, which takes the form

$$\mathcal{L}\{\ln(\boldsymbol{R}/\boldsymbol{S}), \alpha, \beta, \eta_0, \eta_1\} \propto -\frac{1}{2}\sum_{i=1}^{n}(\eta_0+\eta_1 S_i) - \frac{1}{2}\sum_{i=1}^{n}\frac{\left\{\ln\left(\frac{R_i}{S_i}\right) - \ln(\alpha) + \beta S_i\right\}^2}{\exp(\eta_0+\eta_1 S_i)}. \tag{3.10}$$

The Schaefer model is obtained for $\gamma = +1$,

$$\mathcal{L}\{\ln(\boldsymbol{R}/\boldsymbol{S}), \alpha, \beta, \eta_0, \eta_1\} \propto -\frac{1}{2}\sum_{i=1}^{n}(\eta_0+\eta_1 S_i) - \frac{1}{2}\sum_{i=1}^{n}\frac{\left\{\ln\left(\frac{R_i}{S_i}\right) - \ln(\alpha) - \ln(1-\beta S_i)\right\}^2}{\exp(\eta_0+\eta_1 S_i)}. \tag{3.11}$$

I use the AICc score with the point estimate approach for applying model selection instead of approximating the marginal likelihood function. I applied ADMB over the four different models (e.g. $\gamma = \{-2, -1, 0, +1\}$) and select for each population the model with the minimum AICc score. To find a point estimate for $\gamma = +1$, the parameter β should be less than $\{1/\max(S_i)\} = 1$ so as to assert a valid argument for the logarithmic function —both $\boldsymbol{S}$ and $\boldsymbol{R}$ are scaled with their maximum value respectively. The purpose of this analysis is to provide a precise assessment for the non-constant variance (instead of relying only on the Beverton-Holt model) as it fits the data more accurately. I found a prevalence of the Schaefer model for the majority of populations that underscores the importance of using density-dependent models in explaining the S-R relationships (Table 3.5). To better understand and evaluate the impact of the heteroscedasticity coefficient, Figure 3.2 illustrates several plots showing recruits versus relative spawning stock biomass along with the estimated stock-recruit relationship and approximate 95% confidence intervals around this relationship for both heteroscedastic and non-heteroscedastic (i.e. constant variance) models; these plots are for four populations fitted with different shape parameter γ respectively. Recruitment is commonly assumed to have stochastic variability that follows a log-normal distribution from which I derive the expected recruits for each shape parameter γ (see Appendix B for complete derivation). By substituting γ for $-2, -1, 0$ and $+1$ in Equation (B.1), the expected recruits becomes:

$$\mathbb{E}(R) = \frac{\alpha S}{\sqrt{1+2\beta S}}\exp\left(\frac{\exp(\eta_0+\eta_1 S)}{2}\right) \qquad \text{for } \gamma = -2, \tag{3.12}$$

$$\mathbb{E}(R) = \frac{\alpha S}{1+\beta S}\exp\left(\frac{\exp(\eta_0+\eta_1 S)}{2}\right) \qquad \text{for } \gamma = -1, \tag{3.13}$$

$$\mathbb{E}(R) = \alpha S\exp\left(-\beta S + \frac{\exp(\eta_0+\eta_1 S)}{2}\right) \qquad \text{for } \gamma = 0, \tag{3.14}$$

$$\mathbb{E}(R) = \alpha S(1-\beta S)\exp\left(\frac{\exp(\eta_0+\eta_1 S)}{2}\right) \qquad \text{for } \gamma = 1. \tag{3.15}$$

The expected stock-recruitment curves are plotted based on the fitted value of γ,

Table 3.5: Populations fitted with best-fit model parameter γ. The model selection is based on the shape parameter γ corresponding to: $\gamma = -2$ (Cushing-like), $\gamma = -1$ (Beverton-Holt), $\gamma = 0$ (Ricker) and $\gamma = 1$ (Schaefer).

γ	Model	Fitted populations
-2	Cushing	12 out of 90
-1	Beverton-Holt	26 out of 90
0	Ricker	19 out of 90
+1	Schaefer	33 out of 90

the MLE found with ADMB, and the SSB model input (Figure 3.2). I illustrate the heteroscedastic expected recruitment curve with a solid black plot, and the non-heteroscedastic expected recruitment curve with a dotted black plot —obtained by setting $\eta_1 = 0$. I construct the approximate 95% confidence interval for recruitment as follows: sort the SSB population in an ascending order; use Equation (3.3) to generate 10,000 samples for each element; approximate the 95% confidence interval of recruitment estimates for each SSB data point, using the percentile of the sampling distribution. Figure 3.2 shows that the coefficient of heteroscedasticity η_1 has a positive impact in estimating the approximate 95% confidence interval; the coverage of recruits (dashed red plot) is more accurate than the non-heteroscedastic model (grey envelope), and hence its significance in fisheries management. The plots for the 90 S-R populations are illustrated in Appendix C.

3.2.4.1 Measure of Confidence Interval

I used the bias-corrected and accelerated method (BCa) (DiCiccio and Efron, 1996) to form the approximate 95% confidence interval of the density distribution $\hat{\boldsymbol{\eta}}_1^*$. Let $\hat{G}(\hat{\eta_1})$ be the cumulative distribution function (cdf) of bootstrap replications $\hat{\boldsymbol{\eta}_1}^*$,

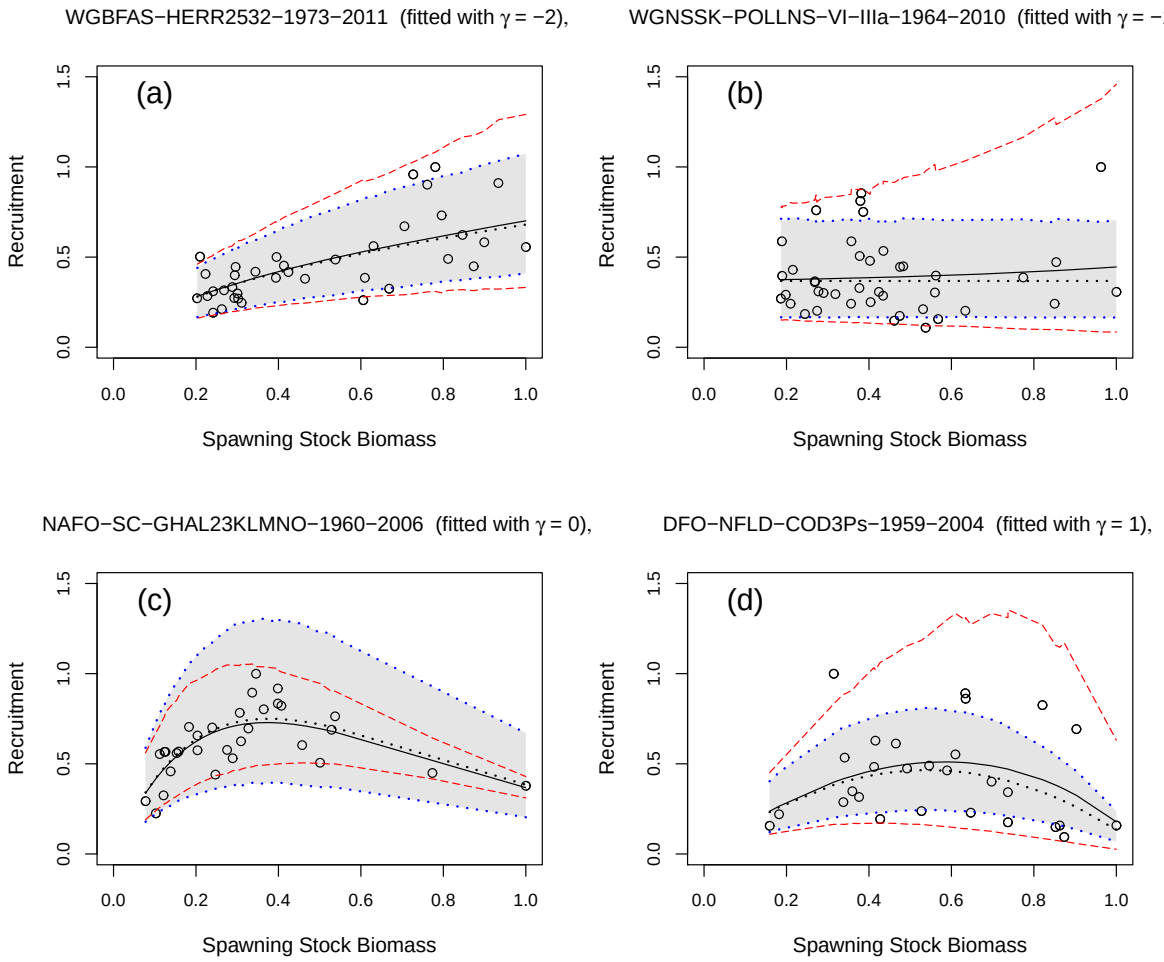
$$\hat{G}(\hat{\eta_1}) = \#(\hat{\boldsymbol{\eta}_1}^* \leq \hat{\eta_1})/B. \tag{3.16}$$

In this case $B = 1000$ replications. By definition the bias-corrected $\kappa/2$ endpoints for the percentile bootstrap confidence interval are calculated as

$$\hat{\eta}_{1BC_a}(\kappa) = \hat{G}^{-1}\left\{\Phi\left(\hat{z}_0 + \frac{\hat{z}_0 + z^{(\kappa)}}{1 - a(\hat{z}_0 + z^{(\kappa)})}\right)\right\}, \tag{3.17}$$

where $\Phi(.)$ is the standard normal cdf. The BCa interval is controlled by two parameters, namely the bias-correction $\hat{z}_0$ and acceleration parameters a. The

Figure 3.2: Expected stock-recruitment curves with approximate 95% confidence intervals fitted with different values of γ. Examples of the Herring, Pollock, Greenland halibut, and Cod families, chosen to illustrate the difference in fit between the heteroscedastic and non-heteroscedastic models. (a) Herring from Eastern Baltic (fitted with $\gamma = -2$), (b) Pollock from IIIa, VI and North Sea (fitted with $\gamma = -1$), (c) Greenland halibut from Labrador Shelf - Grand Banks (fitted with $\gamma = 0$), and (d) Cod from St. Pierre Bank (fitted with $\gamma = +1$). The expected recruit for the non-heteroscedastic model (dotted black plot) and its approximate 95% confidence interval (grey envelop) are compared against the expected recruit for the heteroscedastic model (solid black plot) and its approximated 95% confidence interval (dashed red plot).



bias-correction estimate $\hat{z}_0$ gives the proportion of estimates $\hat{\boldsymbol{\eta}}_1^*$ less than $\hat{\eta}_1$, such that

$$\hat{z}_0 = \Phi^{-1}\left\{\hat{G}(\hat{\eta}_1)\right\} = \Phi^{-1}\left\{\frac{\#(\hat{\boldsymbol{\eta}}_1^* \leq \hat{\eta}_1)}{1000}\right\}, \tag{3.18}$$

where Φ^{-1} is the probit function. DiCiccio and Efron (1996) developed a theory to approximate the confidence interval. The method defines the acceleration parameter a that measures how rapidly standard error changes on a normalized scale, which has an interpretation of skewness. A non-parametric estimate of a can be described as

$$\hat{a} = \frac{1}{6}\frac{\sum_{i=1}^{n} U_i^3}{(\sum_{i=1}^{n} U_i^2)^{3/2}}. \tag{3.19}$$

Then, the jackknife influence function U_i is calculated, as

$$U_i = (n-1)(\hat{\eta}_1 - \hat{\eta}_{1(i)}), \tag{3.20}$$

where $\hat{\eta}_{1(i)}$ is the estimate of η_1 based on the reduced data $\boldsymbol{R}_{(i)} = (R_1, \ldots, R_{i-1}, R_{i+1}, \ldots, R_n)$ and $\boldsymbol{S}_{(i)} = (S_1, \ldots, S_{i-1}, S_{i+1}, \ldots, S_n)$. Therefore, the central 95% BCa interval for η_1 is obtained by

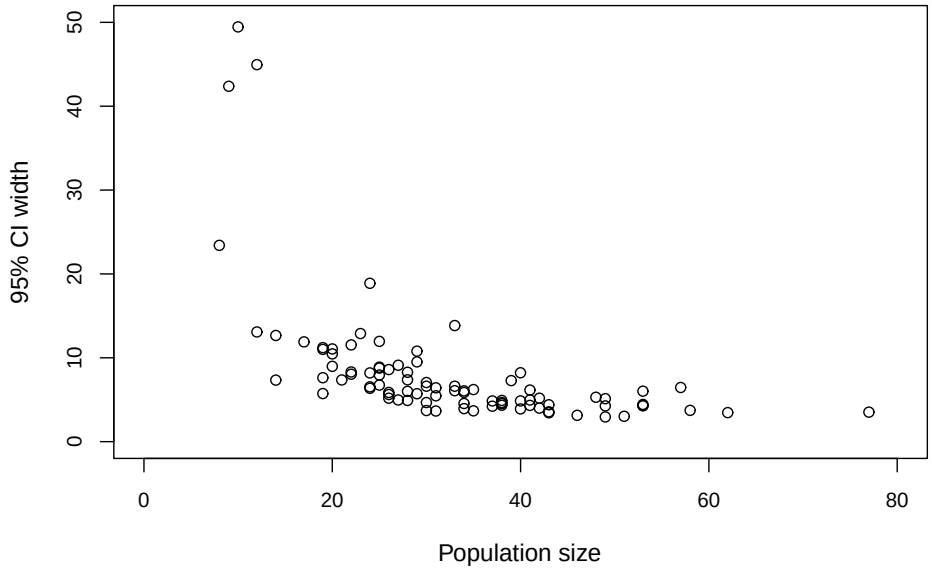
$$\mathrm{CI}_{95\%}(\eta_1) = (\hat{\eta}_{1BC_a}(0.025), \hat{\eta}_{1BC_a}(0.975)). \tag{3.21}$$

Note that the confidence interval for η_1 is mainly influenced by the number of data points found in the population. The more data points there are, the narrower the confidence interval will be. This means that the variability for $\hat{\eta}_1$ is small for large populations, thereby leading to a more reliable fit of the heteroscedastic model than for small population sizes. Figure 3.3 illustrates the approximate 95% BCa confidence interval width versus the population size, applied to all populations.

3.2.4.2 Classification based on the frequentist paradigm

My goal is to investigate whether I could recover accurately the sign of the coefficient of heteroscedasticity. Here, I analyse whether the approximate confidence interval for η_1 denoted as $\mathrm{CI}(\eta_1)$ lies in a region showing a consistent sign with the coefficient η_1. Accordingly, I classify each population in one of three ways: (-1) strong evidence for negative η_1 that is attained when $\mathrm{CI}(\eta_1)$ lies in the negative region; (+1) strong evidence for positive η_1 when $\mathrm{CI}(\eta_1)$ lies in the positive region; and (0) inconclusive evidence for heteroscedasticity.

Figure 3.3: Plot showing the effect of the sample size on the width of the approximated 95% confidence interval. This plot is generated for a Beverton-Holt stock-recruitment model (γ= -1).



3.2.5 Bayesian inference

Bayesian methods offer an alternative to the traditional frequentist method, and may be particularly valuable for populations where there is already some information about the model's parameters. To define a Bayesian analogue of the method of fitting parameters outlined above, I need to specify prior distributions for $\log(\alpha)$, β, η_0 and η_1 to quantify my knowledge before considering the data. Difficulties in Bayesian approaches arise in prior specification such that for a non-informative prior results obtained by using Bayesian methods will be approximately similar to using the frequentist paradigm. However, Bayesian methods may be particularly useful where priors can be specified based on information about similar stocks.

Because there is no prior knowledge for the parameter values, I chose two arbitrary sets of priors to see to what extent the marginal posteriors vary. Firstly, I choose to define a normal prior for $\log(\alpha)$, a wide uniform prior for β covering a region of positive values (to avoid numerical failures), and a symmetrical uniform

prior around the origin for both η_0 and η_1, such that

$$\pi_1\{\log(\alpha)\} = \mathcal{N}(1, 5^2) \tag{3.22}$$
$$\pi_1(\beta) = \mathcal{U}(0, 6000) \tag{3.23}$$
$$\pi_1(\eta_0) = \mathcal{U}(-30, +30) \tag{3.24}$$
$$\pi_1(\eta_1) = \mathcal{U}(-30, +30). \tag{3.25}$$

Secondly, I defined a normal prior for $\log(\alpha)$; a Gamma for β to constrain it to positive values; and a Gaussian prior for both η_0 and η_1, such that

$$\pi_2\{\log(\alpha)\} = \mathcal{N}(1, 5^2) \tag{3.26}$$
$$\pi_2(\beta) = Ga(1, 0.001) \tag{3.27}$$
$$\pi_2(\eta_0) = \mathcal{N}(0, 10^2) \tag{3.28}$$
$$\pi_2(\eta_1) = \mathcal{N}(0, 10^2). \tag{3.29}$$

It is common to assume independent priors for the parameters, such that $\pi(\boldsymbol{\theta}) = \pi\{\log(\alpha)\} \times \pi(\beta) \times \pi(\eta_0) \times \pi(\eta_1)$.
Here I used four different Markov chain Monte Carlo (MCMC) sampling methods: (1) Metropolis within Gibbs, (2) Metropolis Adjusted Langevin Algorithm, (3) Hamiltonian Monte Carlo, and (4) an MCMC package called JAGS (Just Another Gibbs Sampler) (Plummer, 2003); I implemented the first three methods in R and called JAGS, an open-source engine for the BUGS language written in C++, from R via package rjags.

3.2.5.1 Comparison of Markov chain Monte Carlo methods

The implemented MCMC methods in R suffered from slow execution speed compared to JAGS, which is fast and easy to use. These methods are used to sample from the joint posterior distribution $p(\log(\alpha), \beta, \eta_0, \eta_1 | \boldsymbol{R}, \boldsymbol{S})$ so as to estimate the marginal posterior distribution of η_1 given data. For illustrative purposes, I compared convergence of all four methods on a representative dataset (e.g. DFO-QUE-COD3Pn4RS-1964-2007) as described in (Table 3.6). Results show that the HMC sampling method works very well as it provided the largest effective sampling size (ESS) among the other methods; but its weakness lies in its speed: 27 minutes to analyse a single dataset. One could overcome this weakness by implementing it as a multithreaded application in C++. However, JAGS showed itself

to be more efficient as it proved to be fast in execution and convergence over the multiple chains; additionally, it requires fewer parameters for tuning compared to the HMC algorithm, which requires one to tune the number of iterations over the leapfrog integrator and the step size that are set on an adhoc basis. Hence I used JAGS in estimating the marginal posterior distribution of the model parameters. Figure 3.4 illustrates the MCMC samples generated by JAGS for estimating the

Table 3.6: Comparison of diagnostic MCMC methods with four runs and 1000 posterior samples for the model parameters applied to DFO-QUE-COD3Pn4RS-1964-2007 dataset.

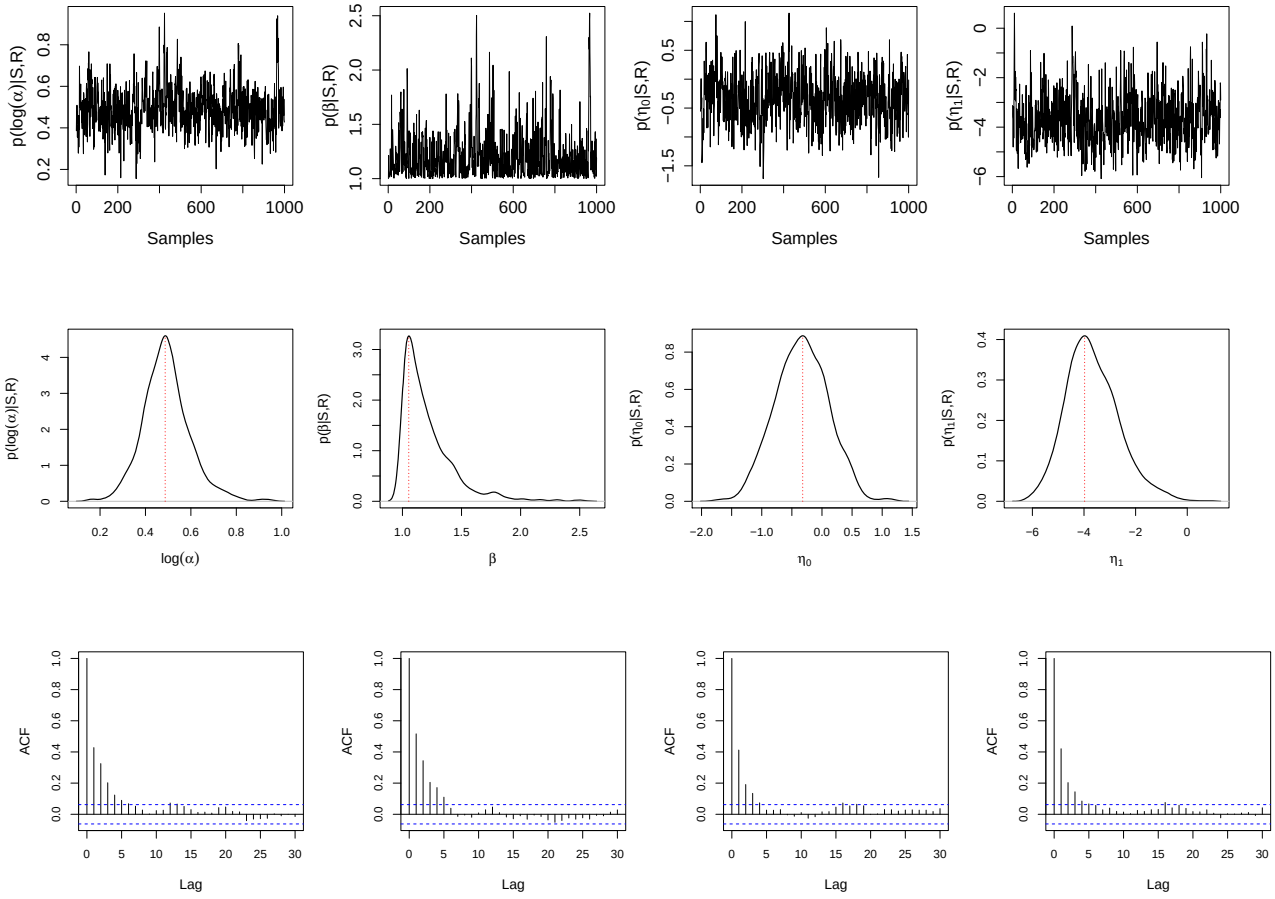
	$\sqrt{\hat{R}}$				ESS				
Method	$\log(\alpha)$	β	η_0	η_1	$\log(\alpha)$	β	η_0	η_1	time(s)
HMC	1.00	1.00	1.02	1.02	250	250	120	250	1643
JAGS	1.02	1.01	1.00	1.02	85.7	87.4	103.9	89.9	15.72
MALA	1.03	1.04	1.04	1.04	98	72	110	120	375
Metropolis-Gibbs	1.14	1.12	1.07	1.07	23	26	43	41	837

marginal posterior distributions for the parameters of interest $(\log(\alpha), \beta, \eta_0, \eta_1)$, ranging from left to right respectively.

3.2.5.2 Convergence criteria

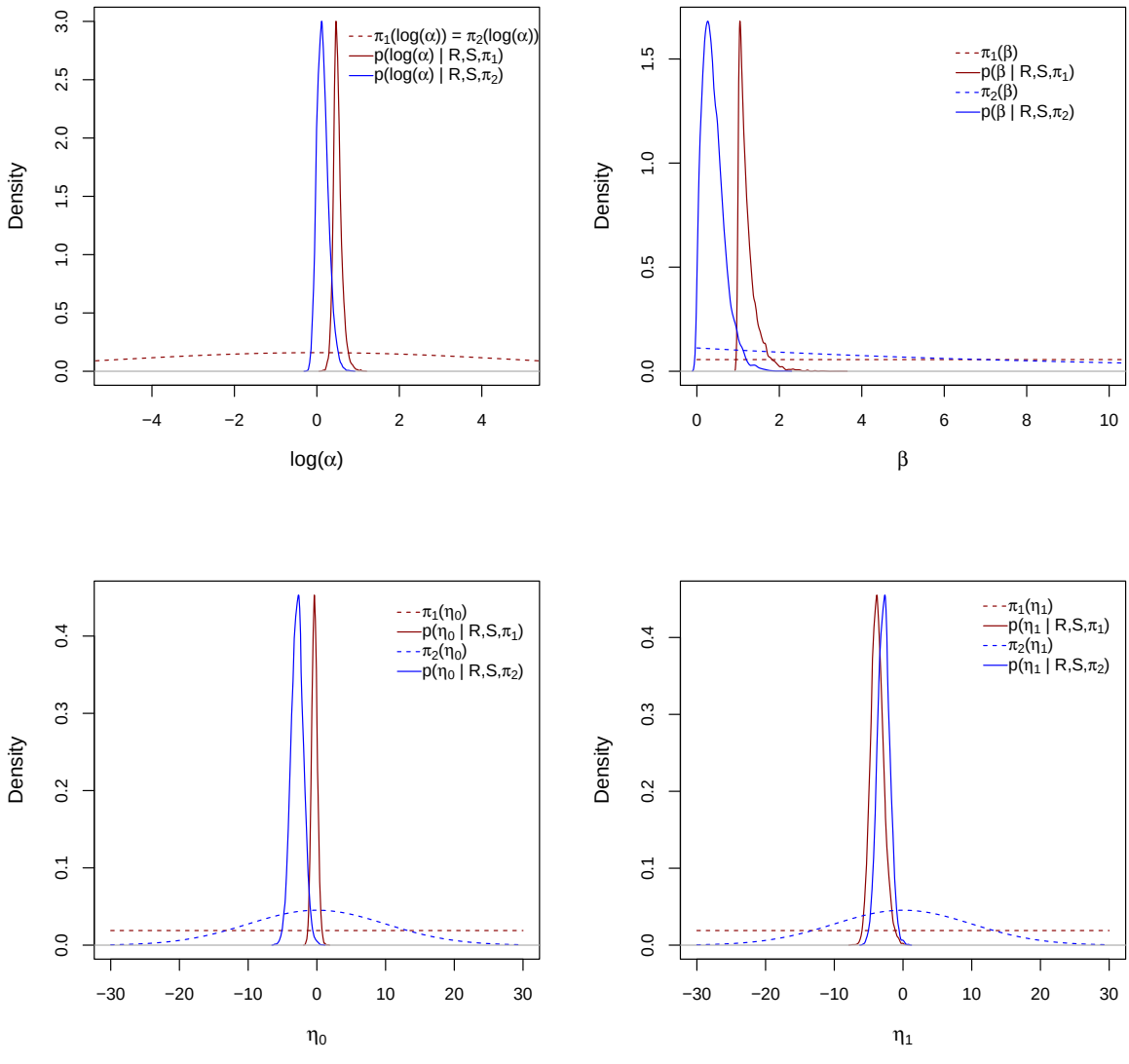
I monitor the approximate convergence of MCMC by using the $\sqrt{\hat{R}}$ statistic provided in the Coda package in R. Gelman (2004) described this statistic as a measure that compares variation between and within simulated sequences until 'within' variation roughly equals 'between' variation, for multiple parallel chains. One can be reasonably confident that convergence is achieved if $\sqrt{\hat{R}} < 1.1$. I simulated four parallel MCMC sequences of 10,000 iterations each after discarding 50,000 samples of each chain, referred to as *burn-in*; these chains are started each from a different initial value and thinned by taking one sample every four samples so as to minimise the autocorrelation between samples. After convergence each simulated sequence is close to the distribution of all other sequences combined together, which all converge to the same posterior distribution. If approximate convergence has not been reached, those populations are identified, then I repeated the approximation by increasing the number of *burn-in* samples (from 1e+5 to 2e+6) and even the number of step size adjustments (from 1e+4 to 1e+5) —tuned by the n.adapt parameter. Figure 3.5 illustrates overlaid plots for the marginal posterior distribution of each parameter of interest $(\log(\alpha), \beta, \eta_0, \eta_1)$, which are shown for priors π_1 and π_2.

Figure 3.4: Graphical displays showing the JAGS sampler output applied onto the DFO-QUE-COD3Pn4RS-1964-2007 dataset. The top row shows trace plots of the marginal distributions of each parameter $(\log(\alpha), \beta, \eta_0, \eta_1)$ and ranging from left to right respectively. The second row shows the empirical marginal posterior distributions of each parameter respectively. The bottom row shows the autocorrelation function for the parameters of interest.



(a) parameter $\boldsymbol{\log(\alpha)}$ (b) parameter $\boldsymbol{\beta}$ (c) parameter $\boldsymbol{\eta_0}$ (d) parameter $\boldsymbol{\eta_1}$

Figure 3.5: Marginal posterior distributions for the parameters produced by the JAGS sampler sampled from priors π_1 and π_2 respectively. Each panel includes four density plots (except for the top left one): two priors (π_1 and π_2)for each parameter and the posteriors corresponding to each of these priors when applied to DFO-QUE-COD3Pn4RS-1964-2007 population.



3.2.5.3 Bayesian sensitivity analysis

The theory of subjective probability enables one to apply a prior distribution in inference to reflect reasonable a priori assumption about parameters. In Bayesian statistics, prior robustness is a real issue for inference; to reduce this concern, one should investigate whether slight changes in the prior distribution cause significant changes in the decision rule. Here I found that the choice between π_1 and π_2 did not influence the resulting Bayesian inference: indicating a reasonable degree of robustness.

3.2.6 Edge Effects Analysis

VPA analysis estimates stock sizes and fishing mortality rates for each year-class (cohort) making up the overall population; the recruitment is estimated as abundance at age 1 and the SSB is estimated by summing up stock sizes of age 2+ in each year respectively (Anderson, 1978). As one goes backward in time, the final age class assumptions and the catch-at-age data totally drive the estimates to become very precise at the beginning of the age group; however, techniques based on the shrinkage to the mean factor —such as XSA—can impose constraints on the last year estimates as well as on the oldest age group (Daskalov, 1998). To account for this kind of estimation error, I analyse the possibility of edge effects in the VPA methods by removing data points from the beginning and end of the time series data. Here I analyse two types of populations. First, I revisit results obtained from model selection by selecting populations having their approximate 95% confidence interval for η_1 lying entirely in the negative region (as shown in Appendix A, Table A.1). Second, I select populations having more than 55 data points so as to check the effect on long time series data. The former selection presents seven full populations for which five were fitted with $\gamma = 1$, one was fitted with $\gamma = 0$, and one was fitted with $\gamma = -2$. Next I truncated two data points at both ends (four points in total) of the seven populations, and five data points at both ends (ten points in total) of the four biggest populations on which I repeated the analysis of testing the reliability of the non-constant variance. If the 95% approximate confidence intervals of both full and truncated datasets are consistent in sign, then I conclude that edge effects in VPA methods are unlikely to influence our results; otherwise I conclude that VPA methods are likely to influence the results.

Table 3.7: Confidence Levels and data classification of the 90 S-R populations for a Beverton-Holt stock-recruitment model; the {-1, 0,+1} coding based on the $\hat{\eta}_1$ distribution indicates the presence of: strong evidence for reliably identifying $\eta_1 < 0$, inconclusive evidence where the sign of η_1 can not be identified, and strong evidence for identifying $\eta_1 > 0$, respectively.

Confidence Level (%)	Coding of $\hat{\eta}_1$ distribution		
	-1	0	$+1$
60	30	43	17
70	26	50	14
80	22	61	7
90	15	72	3
95	11	78	1
99	7	82	1

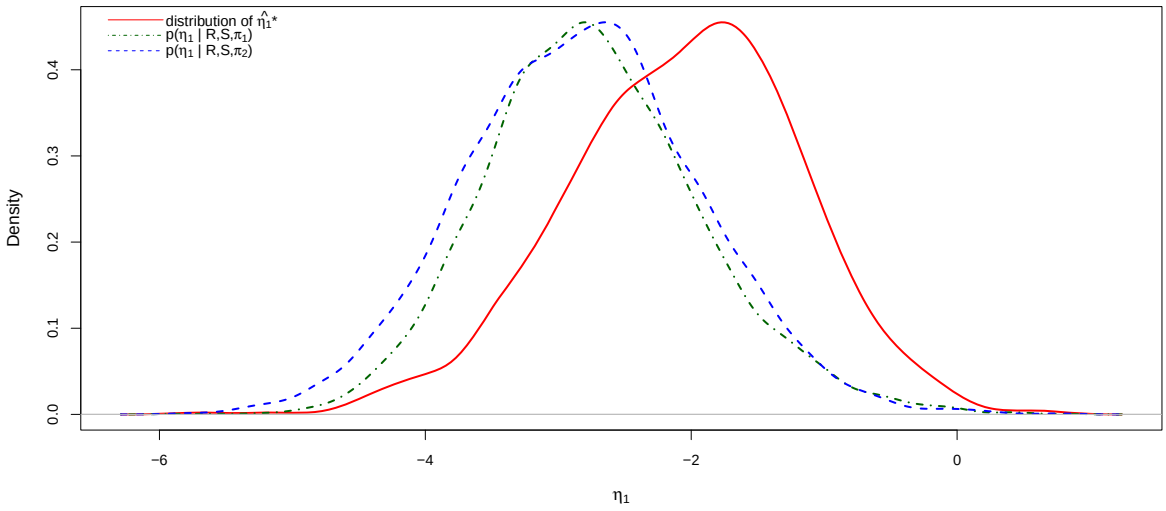
3.3 Results

Statistical analysis based on the frequentist paradigm shows, for a Beverton-Holt model, the existence of seven populations having their approximate 99% confidence interval for η_1 lying entirely in the negative region (Table 3.7). Those seven populations are from six different fish species in six locations, indicating that this classification result is not peculiar to a particular species or location. Standard confidence levels were increased gradually to reflect the sensitivity of the classification labels of the 90 S-R populations to the choice of cut-offs. For low confidence levels I observed many populations classified with label -1, but this classification declined as I increased the confidence level.

Next, I compared results obtained from the frequentist approach to those obtained from Bayesian methods. In the Bayesian framework the credible interval is obtained from the marginal posterior distribution using the equal-tailed credible interval. Figure 3.6 illustrates a comparison between the frequentist and Bayesian inference (for different priors for η_1) applied to a single population, namely DFO-QUE-COD3Pn4RS-1964-2007. Note that in the frequentist method I am using a parametric bootstrap replication of $\hat{\eta}_1$; however, in the Bayesian setting I am estimating the marginal posterior distribution of η_1 given a particular population and a prior. I am in general interested in whether these methods produce the same result or not; the bootstrap estimates (red plot), the posterior distribution with respect to π_1 (dotted-dashed green plot) and the posterior with respect to π_2 (dashed blue plot) produced approximately comparable results in the sense that

their approximate 95% confidence interval and approximate 95% credible intervals were more likely to agree, despite the difference of their shapes. I should also inform the reader that in this figure I used half of the posterior samples (i.e. 5,000 samples) from both posteriors so as to avoid overlap of $p(\eta_1|\mathbf{R},\mathbf{S},\pi_1)$ and $p(\eta_1|\mathbf{R},\mathbf{S},\pi_2)$ in plots. I generalised this comparison —by analysing the output

Figure 3.6: Density plots of: 1,000 parametric bootstrap replications of $\hat{\eta}_1$ (solid red plot); marginal posterior distribution of η_1 with respect to π_1 (dotted-dashed green plot); marginal posterior distribution of η_1 with respect to π_2 (dashed blue plot). The analysis is applied to the DFO-QUE-COD3Pn4RS-1964-2007 population.



of frequentist and Bayesian methods—for all 90 S-R populations, as illustrated in Figure 3.7. The MLE used for bootstrap simulations is represented by an asterisk (*); the black error bars represent the approximate 95% confidence intervals and the square shape denotes the mode of simulated MLEs distribution. The red error bars represent the approximate 95% credible intervals with respect to π_1, and the blue ones represent the approximate 95% credible intervals with respect to π_2. I observed a large approximate 95% confidence interval for the following population numbers: 9, 10, 13, 22, 25, 50, 55, 62 and 63, caused essentially by the small sample sizes: 12, 28, 29, 17, 20, 12, 8, 10, and 9 data points respectively. Moreover, I found for some other populations (20, 50, 53 and 63) different marginal posteriors with respect to the choice of the prior; however, for the remaining populations I found robust posterior inference with respect to the choice

Figure 3.7: Comparison between the frequentist and Bayesian method to inference for a Beverton-Holt model. The black error bars show an approximate 95% BCa confidence interval where the asterisk symbol represents the MLE of η_1 and the square symbol represents the mode of simulated MLEs with bootstrapping. The red error bars and the blue error bars show the approximate 95% credible interval with respect to π_1 and π_2 respectively. The vertical axis represents the η_1 parameter and the horizontal axis represents the sequential population number; ranging from 1 to 30, 31 to 60, and 61 to 90 respectively.

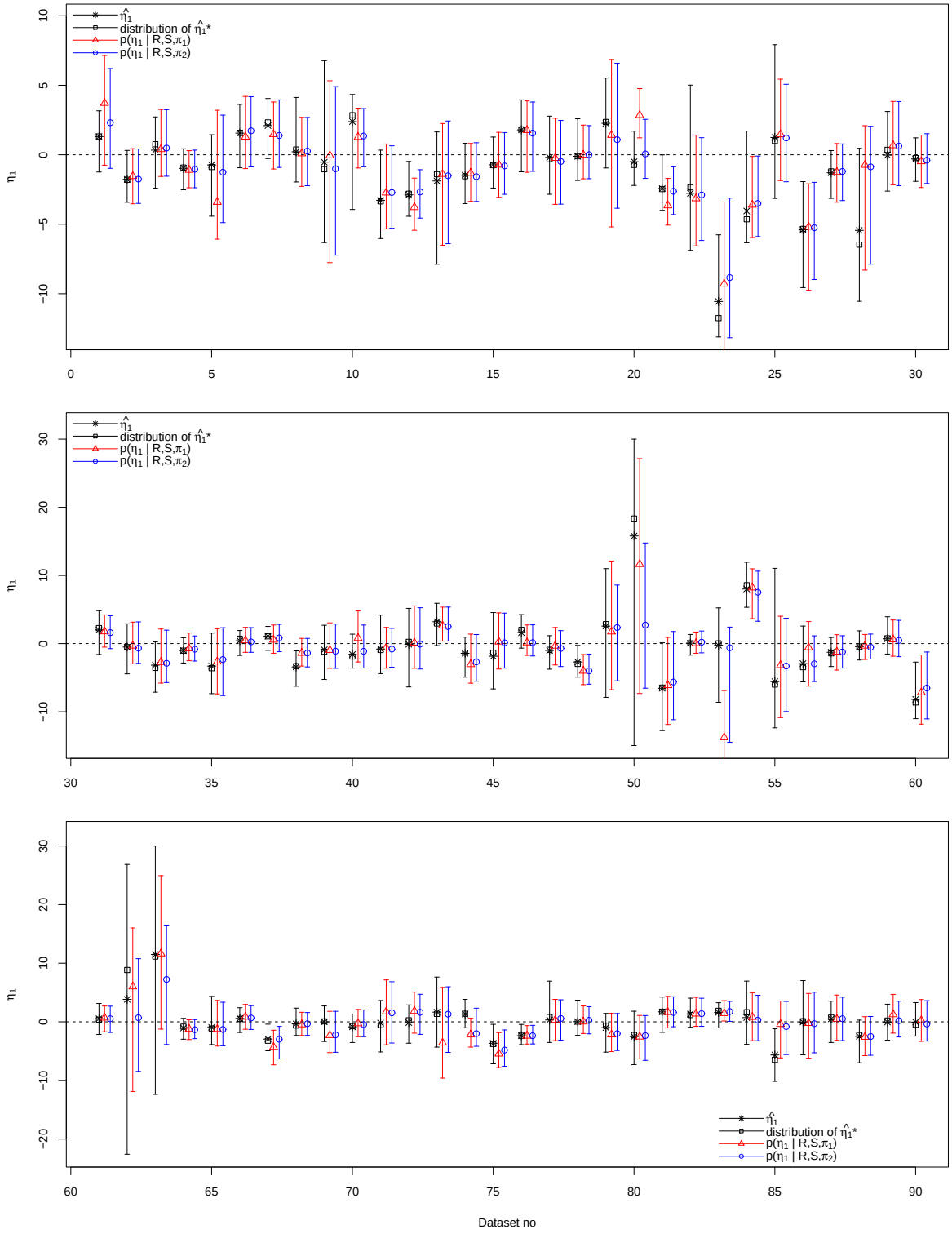


Table 3.8: Comparison between frequentist and Bayesian methods (with π_1 and π_2 priors) for a Beverton-Holt stock-recruitment model for evaluating the reliability of η_1 in survival across the 90 S-R fish populations.

	Coding of η_1 distribution								
	frequentist			Bayesian π_1			Bayesian π_2		
Confidence Level (%)	-1	0	$+1$	-1	0	$+1$	-1	0	$+1$
60	30	43	17	32	43	15	31	47	12
70	26	50	14	29	47	14	27	51	12
80	22	61	7	21	62	7	20	65	5
90	15	72	3	12	73	5	10	77	3
95	11	78	1	11	75	4	10	77	3
99	7	82	1	9	79	2	7	82	1

Table 3.9: Confidence Levels and data classification of the 90 S-R populations using model selection; the {-1, 0,+1} coding based on the $\hat{\eta}_1$ distribution indicate the presence of: strong evidence for reliably identifying $\eta_1 < 0$, inconclusive evidence where the sign of η_1 can not be identified, and a strong evidence for identifying $\eta_1 > 0$, respectively.

	Label		
Confidence Level (%)	-1	0	$+1$
60	31	48	11
70	28	53	9
80	23	62	5
90	16	70	4
95	7	82	1
99	3	87	0

of the prior (i.e. π_1 or π_2). Table 3.8 illustrates a comparison of the estimation error for η_1 assessed by the frequentist and Bayesian approaches when applied to the 90 S-R populations. I observed that both frequentist and Bayesian methods classified approximately the same number of populations, labelled with -1.
To adjust my results, I used the fitted models (derived from model selection) and tested whether I could reliably estimate the sign of η_1 with different confidence levels, as described in Table 3.9. For the case where the confidence level is 95%, I found seven populations labelled with -1, 82 populations labelled with 0 and one population labelled with $+1$. The entire classification list for the 95% confidence level is illustrated in Appendix A (Table A.1).

Finally, I applied the edge effect analysis to populations classified with label -1 and to populations longer than 55 data points (Table A.1). The former revealed an agreement in the classification of six of the seven populations so that I can

Table 3.10: Edge effect analysis applied to populations showing their approximate 95% confidence interval of η_1 lying in the negative region; γ describes the best fitted model, Complete Data describes the CI obtained for the complete population, Truncated Data describes the CI obtained after truncating the population at both ends, and IsComparable indicates whether analysis repeated on truncated population agrees with the original one.

Assessment Id	γ	95% CI		IsComparable
		Complete Data	Truncated Data	
DFO-QUE-COD3Pn4RS-1964-2007	1	[-4.70, -0.22]	[-4.62, -0.72]	Yes
IMARPE-PANCHPERUNC-1963-2004	1	[-5.14, -0.80]	[-5.61 ,-1.25]	Yes
INIDEP-SBWHITARGS-1985-2007	1	[-10.67, -2.27]	[-14.46, -8.47]	Yes
NRIFS-OFLOUNECS-1986-2010	1	[-30.0, -14.2]	[-27.68, -2.50]	Yes
NWWG-HERRIsum-1984-2011	0	[-12.13, -3.53]	[-18.55, -2.62]	Yes
WGBFAS-HERR30-1972-2011	1	[-5.62, -0.16]	[-4.32, -0.09]	Yes
WGNSSK-WHITNS-VIId-IIIa-1989-2010	-2	[-10.23, -0.09]	[-7.60, 3.12]	No

Table 3.11: Edge effect analysis applied to populations with more than 55 data points; γ describes the best fitted model, Complete Data describes the CI obtained for the complete population, Truncated Data describes the CI obtained after truncating the population at both ends, and IsComparable indicates whether analysis repeated on truncated population agrees with the original one.

Assessment Id	γ	95% CI		IsComparable
		Complete Data	Truncated Data	
AFWG-HADNEAR-1947-2010	1	[-3.63, 0.52]	[-3.86, 0.11]	Yes
ICCAT-ATBTUNAEATL-1950-2010	0	[-4.62, 3.09]	[-3.88, 3.49]	Yes
NEFSC-HADGB-1930-2008	1	[-1.45, 2.59]	[-2.12, 2.14]	Yes
WGNSSK-CODNEAR-1943-2010	0	[-4.12, -0.36]	[-4.79, -0.88]	Yes

assert that possible edge effects in the VPA are unlikely to be influencing the analysis for these model inputs (Table 3.10); however, for populations larger than 55 data points I found complete agreement with the original results, indicating that the effect of VPA methods reduces with long time-series populations (Table 3.11).

3.4 Discussion

This study develops, implements and tests methods for identifying non-constant variance (heteroscedasticity) in the spawner-recruit relationship. I found heteroscedastic models tend to fit the S-R model inputs better than constant variance models across the majority of stocks showing a dominance over 78 out of

90 populations (see section 3.2.3) and strong evidence for a negative coefficient of heteroscedasticity in seven cases (Table A.1), including exploited cod, herring and whiting stocks in addition to olive flounder and Peruvian anchoveta. I advocate that the evidence for stochastic regulation in these cases deserves to be taken into account by managers. In contrast, only one stock was identified as having a positive coefficient of heteroscedasticity at the 95% confidence level.

I analysed the estimation error for η_1 (heteroscedasticity parameter) by exploring a class of heterogeneity models —frequentist and Bayesian paradigms—and associated model-fitting algorithms. Under the frequentist method, parameters are viewed as unknown but fixed quantities: consequently the use of inferential procedures were evaluated under repeated sampling of the data. The frequentist method is generally easy to implement; but it encounters difficulties for small population sizes, resulting in a large interval estimation and a loss of statistical significance. In contrast, Bayesian approaches can be appealing for problems of this sort, but difficulties arise in prior specification. Here, I used minimal prior information π_1 and π_2 to obtain the marginal posterior distribution for the parameters of interest; the estimation error for η_1 is obtained by estimating the Bayesian credible intervals using the posterior distribution.

To determine whether I can reliably estimate the sign of η_1, I tested whether the confidence interval lies in a region showing a consistent sign with the coefficient; I found that both frequentist and Bayesian methods led approximately to equivalent inference; but there are some circumstances under which one method outperforms the other, especially when the sample size is below 30 and when the Hessian matrix is not positive definite. The application of model selection reveals a consistent feature across all populations as it selects a model having the best predictive ability among other models; in every case heteroscedastic models fit the data better (i.e. lower AICc score), regardless of the sign of the coefficient of heteroscedasticity. This information is useful in a management context, where knowledge of the coefficient of heteroscedasticity is an important feature in assessing sustainable exploitation regimes (Minto et al., 2008; Burrow et al., 2012). This is illustrated in Appendix A (Table A.1), which broadly labels each population from the set {-1, 0, +1} coding; the value -1 corresponds to stocks where there is good statistical evidence for a negative coefficient of heteroscedasticity (using, in this case, an approximate 95% confidence interval).

To reliably identify a negative coefficient of heteroscedasticity, managers or fisheries scientists using the frequentist methods should check that their chosen confidence interval lies in the negative region; those using the Bayesian framework can consider π_1 or π_2 as a non-informative benchmark prior and check whether their Bayesian credible interval lies in the negative region. I note that Bayesian approaches may be particularly useful where priors can be specified based on information about similar stocks in other locations. To protect this work against false positives or negatives, I recommend fisheries scientists to use both frequentist and Bayesian methods when assessing stocks for heteroscedasticity; if both methods agree then there would be strong evidence that our conclusion is correct; otherwise I should investigate the limitation of each method separately.

Although both frequentist and Bayesian approaches were developed to identify the non-constant variance exhibited in density-dependent models, heteroscedasticity could not be identified for the majority of the datasets no matter which method is used (out of 90 datasets, eight datasets are classified with label -1 and +1, under an approximate 95% confidence interval). The two principal reasons that drive this limited capacity to reliably identify heteroscedasticity are as follows: first, the data are typically rather poorly explained by the best-fitting stock-recruitment relationships due to the inherent noise in the stock-recruitment relationship; second, the time series are not long enough for reliable parameter estimation in most cases (Burrow et al., 2012). Furthermore, it is likely that, in some stocks, the magnitude of any heteroscedasticity is negligible. Nevertheless, this does not diminish the potential importance of heteroscedasticity and its identification, especially in the eight datasets for which I found good evidence of its presence.

I investigated whether there are natural clusterings of stocks with the same heteroscedastic classification; for example, one might hypothesise the same heterscedastic signal of fish stocks of the same (or similar) species in different locations, or alternatively in different stocks at the same location. My preliminary analyses (using approximate 95% confidence levels) indicates no such convenient clusterings. However, further work is needed. For example, classification based on approximate 80% confidence levels reveals a consistent -1 classification for American Plaice, and such patterns may have relevance for sustainable management.

In this analysis, I made two assumptions. First, I discarded ten S-R populations

from the RAM legacy database due to missing data resulting in the analysis of 90 S-R populations of 32 species. I think that this sampling scheme had no bias implication because each population is treated individually and with no effect on the others. Second, I treated all VPA-type assessments as approximately equivalent having first verified that the choice of an assessment had no statistical effect on the sign of coefficient of heteroscedasticity; this is validated by mapping the heteroscedasticity coefficient value against the VPA-type assessment where I found no impact of the VPA-type assessment on the classification method. Additionally, I assessed the possibility of edge effects in the VPA methods. Such effects may be caused by backward-convergence of VPA methods, increased variance of recruitment at the end of the time-series, and shrinkage factors. All these factors may introduce a bias in both SSB and recruitment estimates. This made no difference in the classification of 10 of the 11 populations tested, allowing me to confidently advocate the use of a heteroscedastic model with negative coefficient of heteroscedasticity as a valid management choice in these cases.

My future work will seek to extend the analysis to a more holistic ecosystem-level analysis including external biotic and abiotic factors (i.e. an end-to-end perspective). Besides, I propose combining data from multiple stocks of similar species to better estimate the parameters of the spawner-recruit relationship. To account for heterogeneity, I propose considering a blocking factor and/or within-block correlation in the log-likelihood function across different populations of similar species. If I do not get a blocking effect, meaning that two (or more) populations from the same species have similar variance, then pooling of multiple populations becomes statistically feasible. Alternatively, one can use data from multiple populations to obtain estimates of key parameters for individual populations through a Bayesian hierarchical framework (Gelman, 2004).

Chapter 4

Bayesian Hierarchical Modelling for understanding fish population dynamics and community structures

Hierarchical Bayesian models can be useful for improving estimation of key parameters found in stock-recruitment (S-R) relationships. The presented methodology combines information from various populations to borrow strength across populations so as to estimate improved S-R model parameters across all populations. This allows us to generate one-step-ahead prediction of fish recruitment given stock abundance. I propose four different Bayesian hierarchical models applied on five geographical locations (Celtic Sea, Faroe Plateau, Georges Bank, North-East Arctic, and North Sea) from which I found the non-constant variance model (in the majority of cases) as the best model in predicting fish recruitment values. In addition, I found the grouping of fish species based on water column contributes to better understanding of the dynamics of fish communities compared to the case where the selection is made arbitrarily.
The objectives of this Chapter are the following:

- To examine whether a combined knowledge of fish populations can reduce the uncertainty in key parameters and improve the accuracy of recruitment forecasts.

- To find the best candidate Bayesian hierarchical model that generates accurate predictions of recruitment.
- To find whether the non-constant variance can be reliable to community based fisheries management.

4.1 Introduction

Fisheries stock assessment and management are highly controversial subjects because of imperfect stock sampling and model errors. These factors are the major sources of uncertainty in the population biology of exploited species (Hill et al., 2007), which affect the parameter estimates of the ecological model. The smaller the stock assessment, the more uncertainty is induced in estimating the parameter values. Many scientists have accounted for this problem by introducing some auxiliary data (i.e. tagging data) or prior knowledge that uses survey data and expert judgement to the stock assessments so as to increase the accuracy of the results (Punt et al., 2000; McAllister et al., 1994; McAllister and Ianelli, 1997). Minto et al. (2008) combined data (by species) from different populations using meta-analysis methods to increase the precision of their statistical analysis. Bayesian hierarchical modelling (BHM) offers an alternative way to estimate model parameters especially when dealing with multiple short datasets: one could reasonably expect that the parameters are related to each other such that knowing the parameters of one population would have an influence on the ones from the other population. Many scientists have applied Bayesian analyses using hierarchical models to fisheries stock assessments (Forrest et al., 2010; Michielsens et al., 2008) so as to improve estimates of population characteristics, achieved by combining the data from neighbouring populations. The principle of BHM consists of modelling observable outcomes as conditionally dependent on certain parameters, which themselves are given a probabilistic specification in terms of further parameters, known as hyperparameters (Gelman, 2004). Such representation has the effect of reducing the uncertainty of the biological processes underlying the population dynamics of exploited species (Myers and Mertz, 1998) resulting in improving the quality of stock assessment and developing optimal management strategies (Chen et al., 2003).
In a recent study, Panikian et al. (2015) proposed a future work plan of using a Bayesian hierarchical framework to combine data from multiple stocks so as to

better estimate the heteroscedasticity coefficient. In this work, I use four different stock-recruitment (S-R) relationships: the first is that of Minto et al. (2008), the second is a modified version of Minto's model that incorporates a community-based factor, the third is that of Minto's model but with a constant variance (i.e. $\eta_1 = 0$), and the fourth is the community based model with a constant variance. These models are used within a Bayesian hierarchical framework to inference for random effects (or between-population variance) and predict recruitment values. In general, BHMs provide more accurate statistical estimates of model parameters than the non-Bayesian hierarchical models (non-BHMs) at the expense of some additional parameters that can be regarded as related or connected in some way by the structure of the problem. One can develop a joint posterior distribution of all the unknown quantities so as to reflect the dependencies among the parameters.

To measure the strength of the evidence provided by the data for the different S-R models, I used the deviance information criterion (DIC), the marginal likelihood, and a predictive approach to show comparison among competing models so as to select the one that is best among the candidates. The DIC was proposed by (Spiegelhalter et al., 2002) and is a Bayesian version or generalization of the Akaike information criterion (Akaike, 1973), which trades off a measure of model adequacy against a measure of complexity. Since random effect parameters (see Section 4.2.2) are the focus for predicting results of future recruitment values, deviance based methods such as the DIC metric become appropriate. On the other hand, the marginal likelihood is evaluated by integrating out the random effects and hyperparameters developed under the integrated likelihood (Sinharay and Stern, 2005). The estimation of the marginal likelihood can be achieved in practice by Markov chain Monte Carlo (MCMC) sampling or the Laplace approximation to the completed data likelihood.

The DIC and the marginal likelihood are applied onto training and validation sets constituting 80% of the data; however, the predictive approach based on minimising the root mean square error (RMSE) consists of fitting the model on the training set (60%) and then validating it onto the validation set (20%). However, the last 20% of the data is kept for testing the models.

This analysis is carried out based on data from five different geographical areas: Celtic Sea, Faroe Plateau, Georges Bank, North-East Arctic and North Sea, using data from (Ricard et al., 2012) (www.ramlegacy.org; version 1). The analysis is

also performed on a more global scale: firstly by considering all pelagic populations, then all demersal populations, and finally by combining all populations (demersal and pelagic). The best model is identified for each case respectively from which I predict fish recruitment and evaluate those values against the ones that are assessed by the virtual population analysis (VPA). This approach gives a way for analysing the importance and consistency of the coefficient of heteroscedasticity η_1 within a Bayesian hierarchical framework.
From another perspective, I found an increase in accuracy in predicting fish recruitment by grouping fish populations according to the water column (i.e. pelagic or demersal depths) rather than pooling all fish populations together. For instance, 64% of pelagic fish recruitment are predicted more accurately by restricting the analysis on pelagic populations than pooling all populations together (36% from available test data points); whereas 61% of demersal fish recruitment are predicted more accurately by restricting the analysis on demersal populations than pooling all populations together (39%).

4.2 Materials and methods

4.2.1 The Data

I prune S-R populations collated in the publicly available RAM legacy database (www.ramlegacy.org; version 1) (Ricard et al., 2012) by restricting the analysis only to those estimated by virtual population analysis (VPA) type assessments. The spawning stock biomass (SSB) is measured in tonnes; however, the recruitment is measured in thousands of individuals. The 12 VPA-type assessment methods classified under this category are as follows: VPA, SPA, XSA, FLXSA, ADAPT, NFT-ADAPT, B-ADAPT, SXSA, SPA-ADAPT, NFT-ADAP, ISVPA and hybrid. VPA, also known as cohort analysis, follows cohorts through their whole life, using catch-at-age data and natural mortality to back-calculate what recruitment had to be in order to support the catch (Hilborn and Walters, 1992). In contrast, assessments based on integrated analyses and statistical catch-at-age assessments employ an underlying S-R relationship, so fitting a S-R curve to their time-series is not appropriate. There were 100 S-R populations obtained with VPA-type assessment; but 10 populations had missing data or no data at all, two stocks with no SSB unit (SEFSC-KMACKGM-1992-2001, SEFSC-KMACKSATLC-1981-2001), one

stock with few assessment values containing eight samples (NRIFS-SAURNWPAC-1980-2010), and a single Japanese stock population with very large recruitment numbers (NRIFS-JANCHOPJPN-1978-2009). Accordingly, I restricted our analysis on the remaining 86 fish populations, representing 31 species (see Appendix A, exclude above four stocks from Table A.1) and I scaled all stocks and recruits assessments by $1e^{+6}$ where I preserved the scale across all experiments.

4.2.2 Candidate models for the stock-recruitment relationship

In this work I propose four models $\mathcal{M}_1$, $\mathcal{M}_2$, $\mathcal{M}_3$ and $\mathcal{M}_4$ to analyse the stock and recruitment relationship, such that:

- The first model $\mathcal{M}_1$ is the heteroscedastic stock and recruitment relationship proposed by Minto et al. (2008)

$$\ln\left(\frac{R_{i,j}}{S_{i,j}}\right) \sim_{\text{i.i.d.}} \mathcal{N}(\mu_{i,j}, \sigma^2_{i,j}) \quad \text{where} \quad \mu_{i,j} = \ln(\alpha)_j + \frac{1}{\gamma_j}\ln(1-\gamma_j\beta_j S_{i,j}) \quad \text{and} \tag{4.1}$$

$$\sigma^2_{i,j} = \exp(\eta_{0j} + \eta_{1j} S_{i,j}),$$

  where $R_{i,j}$ and $S_{i,j}$ are the estimated recruitment and spawning biomass for population j in year i respectively; the parameters α_j and β_j measure the productivity and the density-dependent mortality (capacity) in population j, respectively. The density-independent part of the variance is described by η_{0j}, with the density dependent variance described by η_{1j}, known as the heteroscedastic coefficient. For consistent negative values of η_{1j}, the variance varies less with large SSB. The complexity of this model lies in the parameter space of β_j and γ_j due to the constraint imposed on the value of $(1-\gamma_j\beta_j S_{i,j})$ to be positive so as to be computable with the logarithmic function. This creates two disconnected regions such that the parameter β_j is constrained to be positive while γ_j is negative, and vice-versa. To overcome this hurdle, I divided this model into three separate regions:

  1. The parameter γ_j is constrained to be negative and β_j to be positive; in this case the model remains as defined in Equation (4.1).

2. The parameter γ_j is close to 0; in this case I develop a Taylor expansion of $\ln(1-\gamma_j\beta_j S_{i,j})$ up to the third order:

$$\ln\left(\frac{R_{i,j}}{S_{i,j}}\right) \sim_{\text{i.i.d.}} \mathcal{N}(\mu_{i,j},\sigma^2_{i,j}) \quad \text{where} \tag{4.2}$$
$$\mu_{i,j} = \ln(\alpha)_j - \beta_j S_{i,j} - \frac{1}{2}\gamma_j(\beta_j S_{i,j})^2 - \frac{1}{3}\gamma_j^2(\beta_j S_{i,j})^3 \quad \text{and}$$
$$\sigma^2_{i,j} = \exp(\eta_{0j} + \eta_{1j} S_{i,j}),$$

3. The parameter γ_j is constrained to be positive and β_j to be negative; the model remains as defined in Equation (4.1).

- The second model $\mathcal{M}_2$ takes into account fish populations living in a community and subject to competition. I define A as the set of all populations living in a community for which the survival model can be described as

$$\ln\left(\frac{R_{i,j}}{S_{i,j}}\right) \sim_{\text{i.i.d.}} \mathcal{N}(\mu_{i,j},\sigma^2_{i,j}) \quad \text{where} \quad \mu_{i,j} = \ln(\alpha)_j + \frac{1}{\gamma_j}\ln(1-\gamma_j\beta_j C_{i,A}) \quad \text{and} \tag{4.3}$$
$$\sigma^2_{i,j} = \exp(\eta_{0j} + \eta_{1j} C_{i,A}),$$

where $C_{i,A}$ is the normalised community spawning biomass

$$C_{i,A} = \frac{\sum_{j\in A} S_{i,j}}{\max_i(\sum_{j\in A} S_{i,j})}. \tag{4.4}$$

The parameter space of β_j and γ_j is divided similarly into three disconnected regions, as above.

- The third model $\mathcal{M}_3$ is a general Deriso-Schnute (Deriso, 1980; Schnute, 1985) survival model with a constant variance that takes the form

$$\ln\left(\frac{R_{i,j}}{S_{i,j}}\right) \sim_{\text{i.i.d.}} \mathcal{N}(\mu_{i,j},\sigma^2_{i,j}) \quad \text{where} \quad \mu_{i,j} = \ln(\alpha)_j + \frac{1}{\gamma_j}\ln(1-\gamma_j\beta_j S_{i,j}) \quad \text{and} \tag{4.5}$$
$$\sigma^2_{i,j} = \exp(\eta_{0j}).$$

Similarly to $\mathcal{M}_1$ this model is divided into three separate regions:

1. The parameter γ_j is constrained to be negative and β_j to be positive;

in this case the model remains as defined in Equation (4.5).

2. The parameter γ_j is close to 0; in this case I develop a Taylor expansion of $\ln(1-\gamma_j\beta_j S_{i,j})$ up to the third order:

$$\ln\left(\frac{R_{i,j}}{S_{i,j}}\right) \sim_{\text{i.i.d.}} \mathcal{N}(\mu_{i,j},\sigma^2_{i,j}) \quad \text{where} \tag{4.6}$$
$$\mu_{i,j} = \ln(\alpha)_j - \beta_j S_{i,j} - \frac{1}{2}\gamma_j(\beta_j S_{i,j})^2 - \frac{1}{3}\gamma_j^2(\beta S_{i,j})^3 \quad \text{and}$$
$$\sigma^2_{i,j} = \exp(\eta_{0j}),$$

3. The parameter γ_j is constrained to be positive and β_j to be negative; the model remains as defined in Equation (4.5).

- The fourth model $\mathcal{M}_4$ takes into account fish populations living in a community and subject to competition with a constant variance, such that

$$\ln\left(\frac{R_{i,j}}{S_{i,j}}\right) \sim_{\text{i.i.d.}} \mathcal{N}(\mu_{i,j},\sigma^2_{i,j}) \quad \text{where} \quad \mu_{i,j} = \ln(\alpha)_j + \frac{1}{\gamma_j}\ln(1-\gamma_j\beta_j C_{i,A}) \quad \text{and} \tag{4.7}$$
$$\sigma^2_{i,j} = \exp(\eta_{0j}),$$

where $C_{i,A}$ is the normalised community spawning biomass. The parameter space of β_j and γ_j is divided similarly into three disconnected regions, as above.

In brief, the models $\mathcal{M}_1$ and $\mathcal{M}_3$ are based on population-level analysis; however, the models $\mathcal{M}_2$ and $\mathcal{M}_4$ are based on both population and community level analysis, determined by the parameter C.

4.2.3 Current study

I implement four Bayesian hierarchical models for the S-R relationships, as described in equations: (4.1), (4.3), (4.5) and (4.7), to estimate the marginal posterior distribution of random effects and to predict fish recruitment given a value of SSB. This analysis is built on several cases:

- Populations based on geographical locations: I selected the Celtic sea, Faroe Plateau, Georges Bank, North East Arctic and the North Sea that contain three, three, six, five and three populations respectively. All these five

regions contain demersal species except for the North East Arctic area, which contains a single pelagic species.

- Populations based on water column depth: I select 59 populations residing in the demersal water column and 27 populations residing in the pelagic water column as well as all 86 populations.

4.3 Hierarchical Bayesian models

Our computational strategy for Bayesian hierarchical learning follows the general approach that consists of combining the knowledge found in each population to estimate the marginal posterior distribution of the hyperparameters. I employ BHMs that involve three stages for inference; at the first stage, I assume all observations are drawn from a Gaussian distribution. For example, Equation (4.1) shows that the survival variability follows a Gaussian distribution.
At the second stage, the between-population variation is modelled using random effect models. If we believe a priori that all the parameters $(\ln(\alpha)_j, \beta_j, \eta_{0j}, \eta_{1j}, \gamma_j)$ are exchangeable in their joint distribution (de Finetti, 1931), then we can decompose the prior distribution with all the unknown parameters in an i.i.d. mixture form as

$$p(\log(\alpha), \beta, \eta_0, \eta_1, \gamma) = \int p(\log(\alpha)|\boldsymbol{\phi})p(\beta|\boldsymbol{\phi})p(\eta_0|\boldsymbol{\phi})p(\eta_1|\boldsymbol{\phi})p(\gamma|\boldsymbol{\phi})\pi(\boldsymbol{\phi})d\boldsymbol{\phi}, \tag{4.8}$$

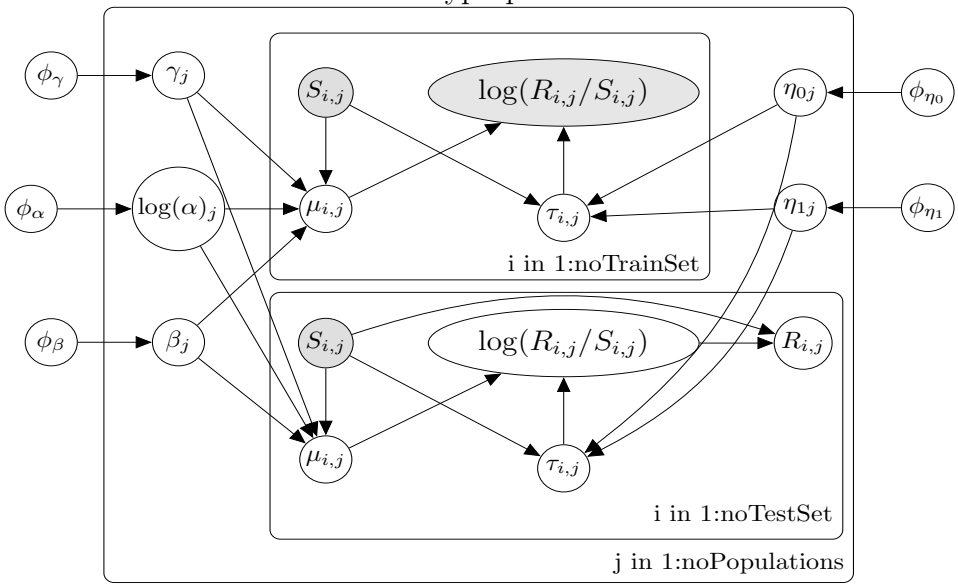
so that the collection of parameters $(\ln(\alpha)_j, \beta_j, \eta_{0j}, \eta_{1j}, \gamma_j)$ are conditionally independent given hyperparameter $\boldsymbol{\phi}$.
At the third stage, a proper hyperprior distribution is set on the hyperparameters such that

$$\boldsymbol{\phi} = (\phi_\alpha, \phi_\beta, \phi_{\eta_0}, \phi_{\eta_1}, \phi_\gamma) \sim_{i.i.d.} \pi(.). \tag{4.9}$$

Parametric choices for $p(.|\boldsymbol{\phi})$ and $\pi(.)$ are described in section 4.3.1. Graphical representations for the BHMs $\mathcal{M}_1$, $\mathcal{M}_2$, $\mathcal{M}_3$ and $\mathcal{M}_4$ are illustrated in Figures 4.1, 4.2, 4.3 and 4.4 respectively. These models are used to make inferences on random effects $(\ln(\alpha)_j, \beta_j, \eta_{0j}, \eta_{1j}, \gamma_j)$ and to predict fish recruitment values using the JAGS software. The analysis takes into consideration the different constraints on γ_j (i.e. $\gamma_j < 0$, $\gamma_j \approx 0$, and $\gamma_j > 0$) and their impact on predictions.
To compare these competing models for finding the best candidate model, I used:

the Deviance information criterion (DIC) (Spiegelhalter et al., 2002); the marginal likelihood estimation via the power posterior method (Friel and Pettitt, 2008); and a predictive approach method (see sections 2.3.3, 4.3.4 and 4.3.5 respectively). Moreover, I compared the BHM against the non BHM (or non-BHM) approach, based on single fish species analysis, to quantify the precision of each method on both inference and prediction.

Figure 4.1: Graphical model illustrating the BHM $\mathcal{M}_1$ for inferring the parameters of Equation (4.1) and forecasting fish recruitment values. The unshaded nodes represent parameters and hyperparameters; the shaded nodes represents the observed data; the rectangular plates denote repetition (i.e. the loop over i and j). For example, $S_{i,j}$ represents the SSB assessment value for population j in year i. The distribution over the hyperpriors is described in section 4.3.1.



4.3.1 Choice for prior and Hyperprior Distributions

A random effects model, also known as multilevel or mixed model, assumes the dataset that we observe is sampled from a larger population; for example, if we collect results from different laboratories, the 'laboratory' might be a random effect. LaMotte (1983) defined that an effect is called random if it is assumed to be a realised value of a random variable. In Bayesian statistics, a prior probability distribution represents our prior belief about model parameters before observing the data. The choice of the random effects and hyperprior distributions for the

Figure 4.2: Graphical model illustrating the BHM $\mathcal{M}_2$ for inferring the parameters of Equation (4.3) and forecasting fish recruitment values. The unshaded nodes represent parameters and hyperparameters; the shaded nodes represents the observed data; the rectangular plates denote repetition (i.e. the loop over i and j). For example, $S_{i,j}$ represents the SSB assessment value for population j in year i. The distribution of the hyperpriors are described in section 4.3.1.

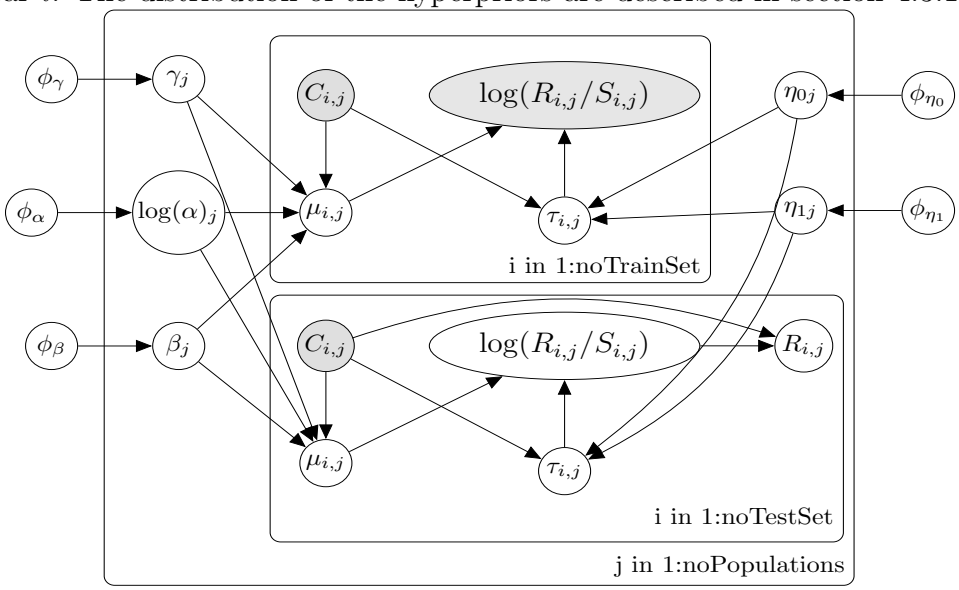


Figure 4.3: Graphical model illustrating the BHM $\mathcal{M}_3$ for inferring the parameters of Equation (4.5) and forecasting fish recruitment values. The unshaded nodes represent parameters and hyperparameters; the shaded nodes represents the observed data; the rectangular plates denote repetition (i.e. the loop over i and j). For example, $S_{i,j}$ represents the SSB assessment value for population j in year i. The distribution of the hyperpriors are described in section 4.3.1.

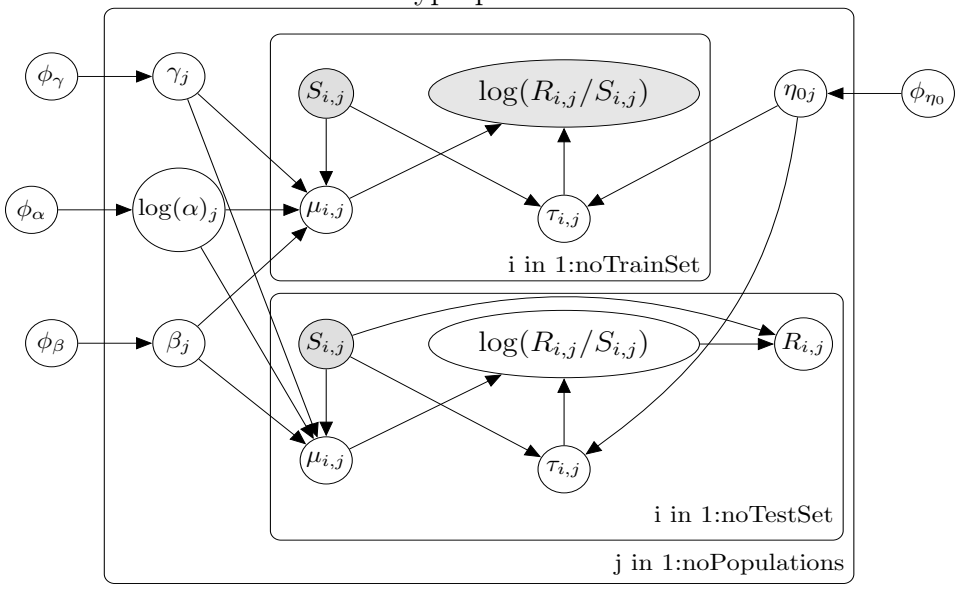
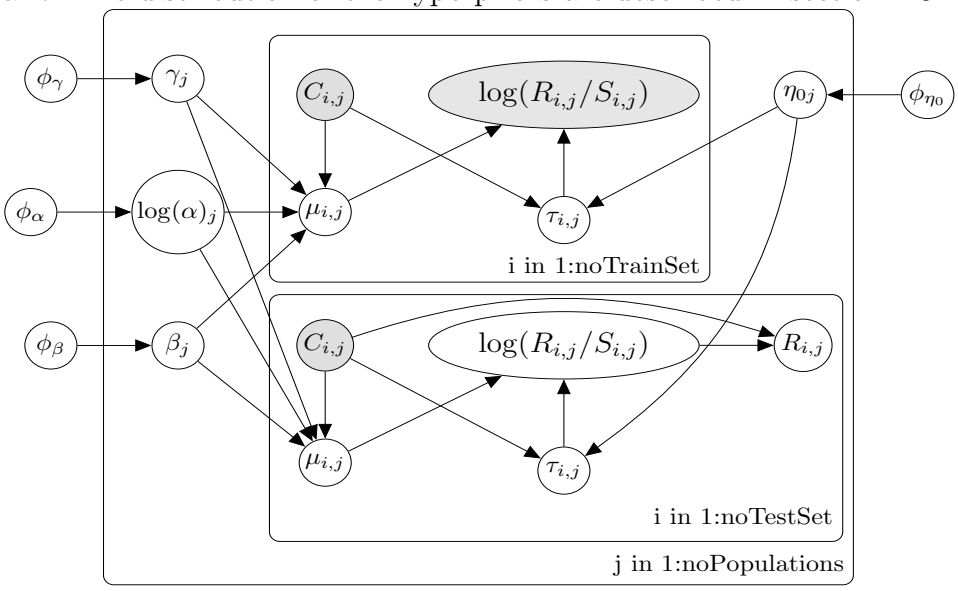


Figure 4.4: Graphical model illustrating the BHM $\mathcal{M}_4$ for inferring the parameters of Equation (4.7) and forecasting fish recruitment values. The unshaded nodes represent parameters and hyperparameters; the shaded nodes represents the observed data; the rectangular plates denote repetition (i.e. the loop over i and j). For example, $S_{i,j}$ represents the SSB assessment value for population j in year i. The distribution of the hyperpriors are described in section 4.3.1.



different ranges of γ_j is based on minimal prior information, such that

$$\gamma_j < 0 : \begin{cases} \log(\alpha)_j & \sim \mathcal{N}(\phi_\alpha, 0.2^2), \quad \phi_\alpha \sim \mathcal{U}(-5, 20) \\ \beta_j & \sim \text{Gamma}(\phi_\beta, 0.01), \quad \phi_\beta \sim \mathcal{U}(1, 300) \\ \eta_{0j} & \sim \mathcal{N}(\phi_{\eta_0}, 0.4^2), \quad \phi_{\eta_0} \sim \mathcal{U}(-20, 20) \\ \eta_{1j} & \sim \mathcal{N}(\phi_{\eta_1}, 1.5^2), \quad \phi_{\eta_1} \sim \mathcal{U}(-20, 20) \\ \gamma_j & \sim \mathcal{N}(\phi_\gamma, 0.2^2), \quad \phi_\gamma \sim \mathcal{U}(-10, -1). \end{cases}$$

The distribution of the productivity parameter $\log(\alpha)_j$ is chosen to cover a wide range such that the uncertainty over its mean value is assumed to vary uniformly between -5 and 20; the distribution of the density-dependence mortality β_j is chosen to be constrained to positive values to which I assign a wide Gamma distribution; the distribution of the variance parameters η_{0j} and η_{1j} are modelled to follow a Gaussian distribution with a large uncertainty on their mean values varying uniformly between -20 and 20 respectively; and the distribution of γ_j is constrained to be negative to which the mean of its Gaussian distribution is set to

vary uniformly between -10 and -1. This constraint is applied for all populations living in the same region, for all $j \in A$.
For the case $\gamma_j \approx 0$, I define:

$$\gamma_j \approx 0 : \begin{cases} \log(\alpha)_j & \sim \mathcal{N}(\phi_\alpha, 0.1^2), \quad \phi_\alpha \sim \mathcal{U}(-5, 20) \\ \beta_j & \sim \mathcal{N}(\phi_\beta, 1.5^2), \quad \phi_\beta \sim \mathcal{U}(-10, 500) \\ \eta_{0j} & \sim \mathcal{N}(\phi_{\eta_0}, 0.4^2), \quad \phi_{\eta_0} \sim \mathcal{U}(-20, 20) \\ \eta_{1j} & \sim \mathcal{N}(\phi_{\eta_1}, 1.5^2), \quad \phi_{\eta_1} \sim \mathcal{U}(-20, 20) \\ \gamma_j & \sim \mathcal{N}(\phi_\gamma, 0.1^2), \quad \phi_\gamma \sim \mathcal{U}(-1, +1), \end{cases}$$

where the distributions of β_j and γ_j are not constrained to any specific region for all populations. Finally, for the case $\gamma_j > 0$, I define:

$$\gamma_j > 0 : \begin{cases} \log(\alpha)_j & \sim \mathcal{N}(\phi_\alpha, 0.2^2), \quad \phi_\alpha \sim \mathcal{U}(-5, 20) \\ \beta_j & \sim \mathcal{N}(\phi_\beta, 0.4^2), \quad \phi_\beta \sim \mathcal{U}(-10, -1.5) \\ \eta_{0j} & \sim \mathcal{N}(\phi_{\eta_0}, 0.4^2), \quad \phi_{\eta_0} \sim \mathcal{U}(-20, 20) \\ \eta_{1j} & \sim \mathcal{N}(\phi_{\eta_1}, 1.5^2), \quad \phi_{\eta_1} \sim \mathcal{U}(-20, 20) \\ \gamma_j & \sim \text{Gamma}(\phi_\gamma, 1), \quad \phi_\gamma \sim \mathcal{U}(0.5, 2), \end{cases}$$

where the distribution of β_j is constrained to be negative and that of γ_j is constrained to be positive for all populations.

4.3.2 Bayesian Inference

Markov chain Monte-Carlo (MCMC) methods are employed to draw successive samples from the joint posterior distribution of all parameters such that each simulated value depends only on the previous simulated value. I implemented the four models $\mathcal{M}_1$, $\mathcal{M}_2$, $\mathcal{M}_3$ and $\mathcal{M}_4$ using JAGS and drew random samples from the posterior distribution of model parameters. JAGS runs a number of alternative sampling algorithms that includes the Metropolis-Hastings algorithm (Metropolis et al., 1953; Hastings, 1970), the Slice sampling (Neal, 2003) and the Adaptive Rejection sampling algorithms (Gilks et al., 1995) to update the full conditionals within a Gibbs sampling scheme. For example, the use of the Metropolis algorithm leads to a type of sampling known as Metropolis-within-Gibbs. These samples can then be used to draw inferences regarding the model

parameters and model fit. The inference runs over four parallel MCMC sequences of 10,000 iterations each after discarding 50,000 samples of each chain, referred to as 'burn-in'. These chains are started randomly at different starting points and thinned by taking one sample from every five samples so as to minimize the autocorrelation between samples. The number of adaptation steps is set to 10,000, that is, the number of adaptive iterations used at the start of the simulation. Convergence of the chains is measured via the Gelman statistic ($\sqrt{\hat{R}}$) provided in the Coda package in R. One can be reasonably confident that convergence is achieved if $\sqrt{\hat{R}} < 1.1$. Once the convergence criterion is met, Markov chains of marginal posterior distributions are generated with respect to the likelihood and prior distributions.

4.3.3 Recruitment Prediction

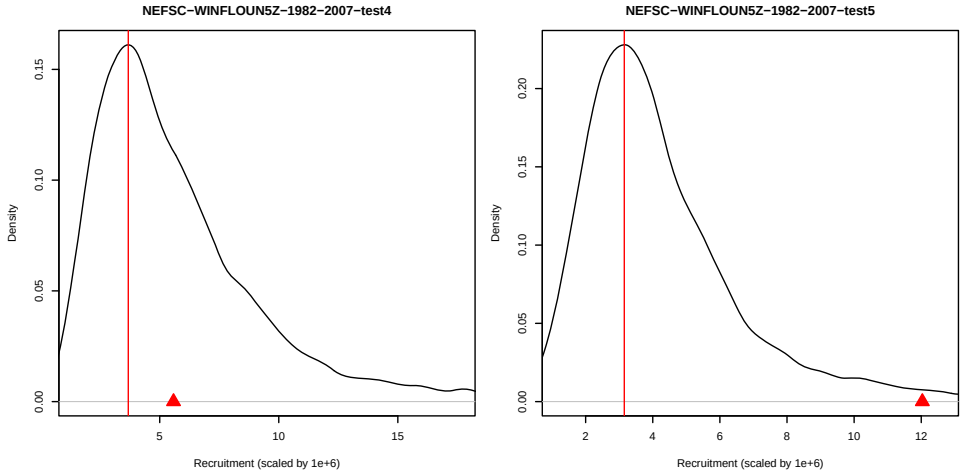
Fish recruitment is predicted using one of possible graphical models (Figure 4.1, 4.2, 4.3 and 4.4) that illustrate how the expected recruitment is predicted from the two sources of data $S_{i,j}$ and $R_{i,j}$. In the course of this analysis, the $S_{i,j}$ is observed in both training and test sets while the $R_{i,j}$ is observed only in the training set. For example, by considering the model illustrated in Figure 4.1, the complete joint posterior distribution of all the parameters and hyperparameters can be described as:

$$\begin{aligned} &p(\mu_{i,j}, \tau_{i,j}, \log(\alpha)_j, \beta_j, \eta_{0j}, \eta_{1j}, \gamma_j, \phi_\gamma, \phi_\alpha, \phi_\beta, \phi_{\eta_0}, \phi_{\eta_1} | S_{i,j}, \log(R_{i,j}/S_{i,j})) \propto \\ &\quad p(\mu_{i,j}|S_{i,j}, \gamma_j, \log(\alpha)_j, \beta_j) \times p(\tau_{i,j}|S_{i,j}, \eta_{0j}, \eta_{1j}) \times p(\log(R_{i,j}/S_{i,j})|\mu_{i,j}, \tau_{i,j}) \times \\ &p(R_{i,j}|\log(R_{i,j}/S_{i,j}), S_{i,j}) \times p(\gamma_j|\phi_\gamma) \times p(\log(\alpha)_j|\phi_\alpha) \times p(\beta_j|\phi_\beta) \times p(\eta_{0j}|\phi_{\eta_0}) \times p(\eta_{1j}|\phi_{\eta_1}). \end{aligned} \tag{4.10}$$

As mentioned in Section 4.3.2 MCMC methods are used to generate a sample from the joint distribution of the parameters. In fact, the sample thus generated is a sample from the joint distribution of all unobserved variables, Equation (4.10), which includes both parameters and test set values of $R_{i,j}$. This sample is then used to estimate the marginal distribution over each test set value of $R_{i,j}$. The expression of the joint distribution can vary based on the selected Bayesian hierarchical model (Figure 4.1, 4.2, 4.3 or 4.4). Appendix E provides the source code of the JAGS model used to predict the marginal posterior distribution of fish recruitment for each test data point. Figure 4.5 illustrates the probability

distributions over possible values of fish recruitment given two test values. The point estimate of fish recruitment is chosen to be the mode of the distributions, which represents a sensible choice for this problem.

Figure 4.5: Density plot for fish recruitment prediction given SSB value. The black curve represents the probability distribution over possible values of predicted recruits where the red vertical line marks the mode of the distribution; however, the red triangle marks the recruitment assessment estimated by VPA. The recruitment axis is scaled by 1e+6.



4.3.4 Marginal likelihood

The power posterior method (Friel and Pettitt, 2008) is a way of estimating the marginal likelihood function for a statistical model by raising the likelihood for the Equation (4.1) (for example) to the power of t:

$$p\{\ln(\boldsymbol{R}/\boldsymbol{S})\,|\alpha,\beta,\eta_0,\eta_1\}^t = \exp\left[t\times\left\{-\frac{n}{2}\ln(2\pi)-\frac{1}{2}\sum_{i=1}^{n}(\eta_0+\eta_1 S_i)-\frac{1}{2}\sum_{i=1}^{n}\frac{\left\{\ln\left(\frac{R_i}{S_i}\right)-\ln(\alpha)-\frac{1}{\gamma}\ln(1-\gamma\beta S_i)\right\}^2}{\exp(\eta_0+\eta_1 S_i)}\right\}\right] \tag{4.11}$$

by taking the logarithm function of both sides, I get

$$t \times \ln[p\{\ln(\boldsymbol{R}/\boldsymbol{S})\,|\alpha,\beta,\eta_0,\eta_1\}] = t \times \left\{ -\frac{n}{2}\ln(2\pi) - \frac{1}{2}\sum_{i=1}^{n}(\eta_0+\eta_1 S_i) - \frac{1}{2}\sum_{i=1}^{n}\frac{\left\{\ln\left(\frac{R_i}{S_i}\right) - \ln(\alpha) - \frac{1}{\gamma}\ln(1-\gamma\beta S_i)\right\}^2}{\exp(\eta_0+\eta_1 S_i)} \right\}. \tag{4.12}$$

A full review of the marginal likelihood method via power posterior is described in Chapter 2, section 2.9. To improve the convergence rate, I started at $t_T = 1$ cutpoint and used the posterior mean parameter values to initiate the chain at the previous cutpoint, t_{T-1}, and so forth. Within each temperature t_s, 4000 samples were collected from the stationary distribution $p_{t_s}\{\boldsymbol{\theta}|\ln(\boldsymbol{R}/\boldsymbol{S})\}$. The implementation of the power posterior method is described in Appendix D.

4.3.5 Predictive approach

I apply the predictive approach on the validation set after dividing the datasets into training, validation and testing sets with proportions of 60%, 20% and 20% respectively. This method consists of evaluating the accuracy of models $\mathcal{M}_1$, $\mathcal{M}_2$, $\mathcal{M}_3$ and $\mathcal{M}_4$ on the validation set and comparing predicted fish recruitment against the VPA estimate using the root mean square error (RMSE) evaluation metric. Our interest is then to select the model that gets the best RMSE most times.

4.3.6 Dataset Split

The dataset is split according to the different evaluation metric techniques. The split for the DIC and MLL techniques consists of dividing the data into training and test sets with proportions of 80% and 20% respectively. The training set is used to fit the models, but the test set is used for testing the accuracy of models. However, the split for the predictive approach consists of dividing the dataset into training, validation and test sets with proportions of 60%, 20% and 20% respectively. The validation set is used to validate the models before applying them on the test set.

4.4 Results

In accordance with the above plan, first I verified the assumption that uncertainty for a BHM is smaller than the non-BHM. Table 4.1 compares the posterior probabilities of both methods and shows that for the same set of priors we achieved a reduced uncertainty or a smaller credible interval width (CI width) for estimates of model parameters when using a BHM compared to a non-BHM. Additionally, the three evaluation metrics (DIC, MLL and predictive approach) indicate that a BHM fits the data better and provides a better predictive accuracy than a non-BHM. Then I employed the model $\mathcal{M}_1$ (with $\gamma \approx 0$) to compare the predictive accuracy of BHM versus non-BHM using the RMSE metric (Table 4.2). This comparison is applied on the North-East Arctic area that comprises 48 test data points from which 36 test points (or 75% of cases) are best predicted with the BHM and the remaining 12 test points (or 25% of cases) are best predicted with the non-BHM, which gives a dominance of BHM over the non-BHM. There is a clear pattern that those remaining 12 test points (or 25% of cases) are related to AFWG-GHALNEAR and WGNSSK-CODCOASTNOR populations (Table 4.2) for which a non-BHM is more suitable than a BHM.

After verifying that the prediction accuracy of BHM outperforms the non-BHM, I employed the BHM to assess the predictive accuracy of models ($\mathcal{M}_1$, $\mathcal{M}_2$, $\mathcal{M}_3$ and $\mathcal{M}_4$) over eight different cases (Celtic sea, Faroe Plateau, Georges Bank, North-East Arctic, north sea, Pelagic, Demersal and All populations) using the DIC, MLL and the predictive approaches. I found that the three metrics have generally disagreed upon each case for the choice of best predictive model. In this work, I followed the general approach in statistics that consists of keeping a separate set for testing purposes so as to identify the best evaluation metric. DIC was slightly better than the other evaluation metrics in depicting the accurate predictive model in three out of eight cases (Tables 4.3 and 4.4), which makes it suitable for this problem. The consistency in sign of η_1 describes its degree of reliability such that a consistent negative sign of η_1 indicates that the variance attenuates with large SSB densities, but a consistent positive sign of η_1 indicates that the variance increases with large SSB densities; otherwise, when the sign becomes inconsistent (the density distribution lies between negative and positive regions) η_1 becomes unreliable.

The test set is used to check the results obtained by the validation set (Table 4.3 and 4.4). It turns out that the test set predictions differ in four out of eight

Table 4.1: Descriptive comparison of posterior distributions resulting from the Bayesian hierarchical model $\mathcal{M}_1$ to those from the equivalent non-hierarchical model applied to fisheries located in North-East Arctic with $\gamma < 0$. The fitting of the two models is assessed with the DIC, MLL and Predictive approach metrics; the smaller the DIC values indicate a better fitting model; however, the larger the MLL values (closer to zero) indicate a better fitting model; and the larger the predictive approach indicate a better fitting model.

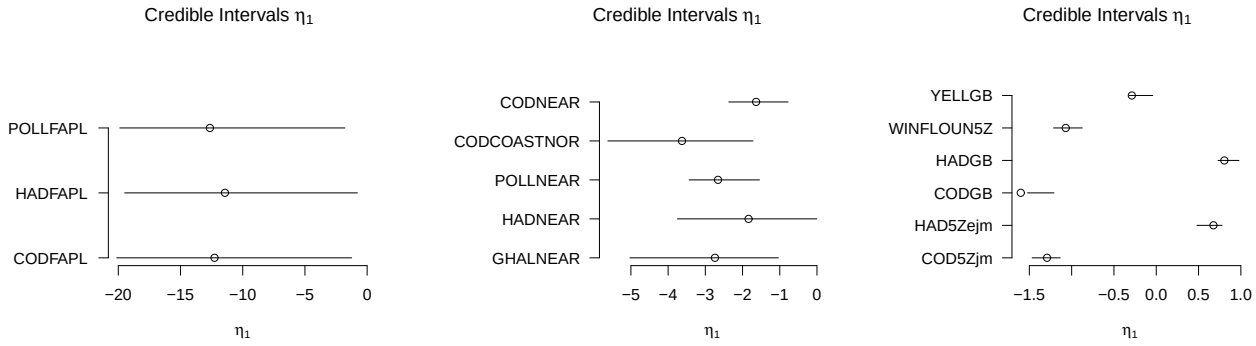
		Hierarchical model Posterior			Non-hierarchical model Posterior		
Population	Parameters	Mode	(95% interval)	CI width	Mode	(95% interval)	CI width
CODNEAR	$\log(\alpha)$	8.53	(7.90, 9.47)	1.57	8.69	(7.95, 9.43)	**1.48**
	β	265.96	(129.29, 964.35)	**835.06**	414.48	(175.87, 1022.39)	846.52
	η_0	-1.72	(-2.21, -1.12)	**1.09**	-2.6	(-3.35, -1.50)	1.85
	η_1	-2.68	(-6.20, 1.25)	**7.45**	4.29	(-9.03, 19.02)	28.05
	γ	-1.26	(-1.68, -1.00)	**0.68**	-1.32	(-1.89, -1.07)	0.82
CODCOASTNOR	$\log(\alpha)$	8.66	(7.96, 9.70)	**1.74**	7.77	(7.42, 9.61)	2.19
	β	42.51	(17.30, 411.13)	**393.83**	392.34	(162.28, 1024.30)	862.02
	η_0	0.17	(-0.31, 0.66)	**0.97**	1.01	(0.03, 2.25)	2.22
	η_1	-1.55	(-4.54, 2.27)	**6.81**	-7.93	(-16.59, 1.56)	18.15
	γ	-1.46	(-1.99, -1.00)	**0.99**	-8.71	(-9.82, -1.81)	8.01
POLLNEAR	$\log(\alpha)$	8.65	(8.02, 9.76)	**1.74**	10.5	(9.20, 11.58)	2.38
	β	38.49	(17.14, 288.18)	**271.04**	401	(159.82, 1033.48)	873.66
	η_0	-0.88	(-1.54, -0.24)	**1.3**	-1.05	(-1.81, -0.05)	1.76
	η_1	-1.99	(-3.72, -0.11)	**3.61**	-1.93	(-4.08, 0.61)	4.69
	γ	-1.31	(-1.79, -0.97)	0.82	-1.16	(-1.82, -1.02)	**0.8**
HADNEAR	$\log(\alpha)$	8.56	(7.91, 9.57)	**1.66**	6.58	(6.31, 9.62)	3.31
	β	179.62	(81.24, 729.59)	**648.35**	389.06	(159.86, 1025.47)	865.61
	η_0	-1.34	(-2.00, -0.65)	**1.35**	-1.54	(-3.73, 0.76)	4.49
	η_1	-3.37	(-6.94, 0.32)	**7.26**	-3.18	(-19.50, 10.75)	30.25
	γ	-1.3	(-1.80, -0.95)	**0.85**	-1.63	(-9.60, -1.15)	8.45
GHALNEAR	$\log(\alpha)$	8.61	(8.06, 9.81)	**1.75**	10.42	(8.55, 11.77)	3.22
	β	7.79	(3.49, 105.12)	**101.63**	397.27	(168.27, 994.27)	826
	η_0	-0.56	(-1.06, 0.04)	**1.1**	-0.46	(-1.11, 0.35)	1.46
	η_1	-1.4	(-2.61, 0.06)	**2.67**	-1.55	(-2.97, 0.50)	3.47
	γ	-1.45	(-1.94, -0.96)	**0.98**	-1.63	(-5.46, -1.24)	4.22
—hyperparameters—							
	ϕ_α	8.56	(8.04, 9.62)	—	—	—	—
	ϕ_β	1.17	(1.02, 4.07)	—	—	—	—
	ϕ_{η_0}	-0.88	(-1.35, -0.34)	—	—	—	—
	ϕ_{η_1}	-1.88	(-4.37, 0.33)	—	—	—	—
	ϕ_γ	-1.36	(-1.77, -1.06)	—	—	—	—
—model comparison—							
	DIC		**343**			344.4	
	MLL		**-152.19**			-154.23	
	Predictive approach		29			**19**	

Table 4.2: Comparison between BHM versus non-BHM using the model $\mathcal{M}_1$, which is applied on the test set of the North-East Arctic area based on the RMSE metric.

	RMSE			RMSE	
Test point	BHM	non-BHM	Test point	BHM	non-BHM
AFWG-GHALNEAR-1960-2010-test1	11.22	**9.33**	AFWG-POLLNEAR-1957-2011-test4	**202.68**	263.44
AFWG-GHALNEAR-1960-2010-test2	10.80	**8.55**	AFWG-POLLNEAR-1957-2011-test5	**181.59**	294.07
AFWG-GHALNEAR-1960-2010-test3	12.40	**8.66**	AFWG-POLLNEAR-1957-2011-test6	**128.35**	271.38
AFWG-GHALNEAR-1960-2010-test4	15.84	**11.10**	AFWG-POLLNEAR-1957-2011-test7	**88.98**	NA
AFWG-GHALNEAR-1960-2010-test5	18.51	**12.15**	AFWG-POLLNEAR-1957-2011-test8	**87.91**	NA
AFWG-GHALNEAR-1960-2010-test6	21.11	**12.99**	AFWG-POLLNEAR-1957-2011-test9	**79.37**	NA
AFWG-GHALNEAR-1960-2010-test7	21.12	**14.08**	AFWG-POLLNEAR-1957-2011-test10	**95.38**	248.52
AFWG-GHALNEAR-1960-2010-test8	**23.21**	23.28	WGNSSK-CODCOASTNOR-1982-2010-test1	**55.35**	56.52
AFWG-GHALNEAR-1960-2010-test9	20.64	**13.72**	WGNSSK-CODCOASTNOR-1982-2010-test2	50.02	**32.51**
AFWG-HADNEAR-1947-2010-test1	**356.66**	15723.44	WGNSSK-CODCOASTNOR-1982-2010-test3	55.44	**38.63**
AFWG-HADNEAR-1947-2010-test2	**452.12**	8218.28	WGNSSK-CODCOASTNOR-1982-2010-test4	48.44	**44.43**
AFWG-HADNEAR-1947-2010-test3	**353.41**	5894.39	WGNSSK-CODCOASTNOR-1982-2010-test5	61.20	**48.65**
AFWG-HADNEAR-1947-2010-test4	**360.22**	8459.52	WGNSSK-CODNEAR-1943-2010-test1	**541.09**	1464.18
AFWG-HADNEAR-1947-2010-test5	**252.87**	9602.55	WGNSSK-CODNEAR-1943-2010-test2	**526.36**	1214.43
AFWG-HADNEAR-1947-2010-test6	**298.35**	10860.39	WGNSSK-CODNEAR-1943-2010-test3	**470.64**	1111.73
AFWG-HADNEAR-1947-2010-test7	**328.20**	6225.47	WGNSSK-CODNEAR-1943-2010-test4	**437.33**	868.94
AFWG-HADNEAR-1947-2010-test8	**363.03**	15009.30	WGNSSK-CODNEAR-1943-2010-test5	**455.36**	774.94
AFWG-HADNEAR-1947-2010-test9	**612.94**	6650.72	WGNSSK-CODNEAR-1943-2010-test6	**519.66**	1102.71
AFWG-HADNEAR-1947-2010-test10	**979.37**	16336.43	WGNSSK-CODNEAR-1943-2010-test7	**480.46**	1220.29
AFWG-HADNEAR-1947-2010-test11	**863.26**	4707.23	WGNSSK-CODNEAR-1943-2010-test8	**531.61**	1283.44
AFWG-HADNEAR-1947-2010-test12	**372.98**	12075.20	WGNSSK-CODNEAR-1943-2010-test9	**537.89**	1335.49
AFWG-POLLNEAR-1957-2011-test1	**170.44**	219.24	WGNSSK-CODNEAR-1943-2010-test10	**508.04**	1151.85
AFWG-POLLNEAR-1957-2011-test2	**153.78**	243.92	WGNSSK-CODNEAR-1943-2010-test11	**570.06**	1352.08
AFWG-POLLNEAR-1957-2011-test3	**137.53**	248.15	WGNSSK-CODNEAR-1943-2010-test12	**694.27**	1495.03

cases from those of the validation set. Moreover, the test set indicates that the heteroscedastic model ($\eta_1 \neq 0$) is found in general to provide a better prediction of fish recruitment value than the non-heteroscedastic models ($\eta_1 = 0$). In five out of eight cases (results showed in Table 4.4) the heteroscedastic model is found to provide a higher accuracy than the non-heteroscedastic model, as assessed by the RMSE metric. The testing set indicates $\mathcal{M}_1$ as the best model for predicting fish recruitment over the majority of cases; however, recruitment variability in Georges Bank —encompasses six demersal fish populations—is best described with the community factor model $\mathcal{M}_2$. By analysing the reliability of the heteroscedasticity parameter using $\mathcal{M}_1$ and $\mathcal{M}_2$ over the three areas (Faroe plateau, Georges Bank and North-East Arctic) where recruitment variability is best described with η_1, I found a consistency in sign of η_1 over an approximate 95% credible interval just in Faroe plateau; however, for the other two areas North-East Arctic and Georges Bank the consistency was reached when the confidence level is lowered down to 70% and 12% respectively (Figure 4.6).

Figure 4.6: Credible interval of η_1 approximated with different confidence levels.



(a) Faroe plateau, 95% CI. (b) North-East Arctic, 70% CI. (c) Georges bank, 12% CI.

Table 4.3: Descriptive comparison of model $\mathcal{M}_1$, $\mathcal{M}_2$, $\mathcal{M}_3$ and $\mathcal{M}_4$ assessed by the means of marginal log-likelihood (MLL), deviance information criterion (DIC) and predictive approach methods. The larger the MLL values (closer to zero) indicate a better fitting model. Note that smaller DIC values indicate a better fitting model; but a larger value for the predictive approach value indicates a better fit.

		$\mathcal{M}_1$			$\mathcal{M}_2$			$\mathcal{M}_3$ ($\eta_1 = 0$)			$\mathcal{M}_4$ ($\eta_1 = 0$)		
Area	method	$\gamma < 0$	$\gamma \approx 0$	$\gamma > 0$	$\gamma < 0$	$\gamma \approx 0$	$\gamma > 0$	$\gamma < 0$	$\gamma \approx 0$	$\gamma > 0$	$\gamma < 0$	$\gamma \approx 0$	$\gamma > 0$
Celtic Sea	MLL	-49.69	-49.64	-50.26	-50.35	**-49.13**	-50.06	-49.69	-49.65	-50.26	-50.34	-49.24	-50.05
(3 dem. pop.)	DIC	149.5	237.9	208.2	213	205.4	188.6	**148.2**	237.8	207.8	207	199.8	188.1
	Predictive approach	**7**	0	0	0	2	1	**7**	0	0	2	1	2
Faroe Plateau	MLL	-94.13	**-92.57**	-94.6	-94.31	-93.28	-94.67	-94.13	-92.62	-94.61	-94.33	-92.92	-94.67
(3 dem. pop.)	DIC	269.6	**245.4**	313	289.1	277.5	319.5	257.1	248.2	310.6	287.3	275.9	317.2
	Predictive approach	1	5	3	**12**	4	0	0	6	0	0	0	0
Georges Bank	MLL	-135.72	-134.76	-136.21	-135.61	-134.60	-136.35	-135.71	-134.75	-136.24	-135.59	**-134.04**	-136.36
(6 dem. pop.)	DIC	491.5	584.5	538.7	479.5	524.3	549	487.9	582.3	540	**473.9**	518.7	547.5
	Predictive approach	7	0	0	7	0	4	2	0	0	**13**	8	4
North-East Arctic	MLL	-152.19	-150.45	-153.52	-152.89	**-150.19**	-153.40	-152.26	-150.56	-153.55	-153.06	-151.05	-153.42
(4 dem. + 1 pel. pop.)	DIC	**343**	374.8	469	409.7	411.4	454.3	346.1	381.8	471	422.7	426.5	454.5
	Predictive approach	6	2	0	5	3	5	**12**	9	0	3	0	3
North Sea	MLL	-79.74	-78.88	-80.22	-80	-78.93	-80.07	-79.73	-78.88	-80.21	-80	**-78.87**	-80.07
(3 dem. pop.)	DIC	226.2	256.5	273.4	253.5	257.5	256.1	**224.9**	254	272.5	251.3	253.1	256.5
	Predictive approach	6	0	1	1	1	2	**10**	3	1	3	0	0
Pelagic	MLL	-447.54	-438.20	-448.09	-441.13	**-435.67**	-447.16	-449.70	-438.48	-448.32	-442.13	-436.63	-448.27
(27 pop.)	DIC	1691	2373	2557	1887	2214	2449	**1639**	2388	2582	1923	2221	2546
	Predictive approach	17	7	5	**73**	35	26	6	2	0	1	7	5
Demersal	MLL	-934.60	-924.03	-940.40	-935.41	-924.80	-936.89	-934.71	-924.38	-940.76	-936.23	**-918.9**	-937.66
(59 pop.)	DIC	**3250**	3935	3912	3425	3411	3569	3285	3954	3946	3495	3466	3629
	Predictive approach	**80**	11	9	50	45	66	27	12	1	25	55	11
All Populations	MLL	-1382.44	-1366.44	-1391.13	-1379.32	-1374.38	-1385.44	-1382.54	-1367.14	-1391.45	-1380.54	**-1357.99**	-1386.87
(86 pop.)	DIC	3964	4325	4432	**3610**	3783	4136	5254	6731	6791	5883	5915	6263
	Predictive approach	**164**	9	9	84	108	97	14	4	1	35	41	10

Table 4.4: Predictive approach for evaluating models $\mathcal{M}_1$, $\mathcal{M}_2$, $\mathcal{M}_3$ and $\mathcal{M}_4$ applied to the testing set. The numbers scored in a model represent the cases for which this particular model is found to minimise the RMSE.

	$\mathcal{M}_1$			$\mathcal{M}_2$			$\mathcal{M}_3$ ($\eta_1 = 0$)			$\mathcal{M}_4$ ($\eta_1 = 0$)		
Predictive approach	$\gamma < 0$	$\gamma \approx 0$	$\gamma > 0$	$\gamma < 0$	$\gamma \approx 0$	$\gamma > 0$	$\gamma < 0$	$\gamma \approx 0$	$\gamma > 0$	$\gamma < 0$	$\gamma \approx 0$	$\gamma > 0$
Celtic Sea	4	0	0	0	1	3	**7**	0	0	0	1	6
Faroe Plateau	0	**12**	0	3	6	0	1	4	0	1	2	2
Georges Bank	11	0	0	**14**	1	1	3	2	0	2	5	6
North-East Arctic	7	**19**	0	5	0	0	11	2	1	1	0	2
North Sea	4	0	0	2	0	0	8	1	2	1	1	**9**
Pelagic	**86**	13	1	8	10	15	0	2	4	0	20	25
Demersal	**87**	32	15	40	49	28	26	6	5	8	47	49
All Populations	86	44	16	37	49	87	27	7	3	35	83	**102**

Finally, I used the test set to measure the effect of grouping fish populations according to the water column (i.e. pelagic or demersal depths) versus the case of pooling all fish populations together on the accuracy of predicting fish recruitment. First, I fit $\mathcal{M}_1$ ($\gamma < 0$) to pelagic populations (i.e. 27 populations) and $\mathcal{M}_4$ ($\gamma > 0$) to the case of all populations (i.e. 86 populations). The predicted recruitment values derived from each analysis are compared against the VPA assessments: showing better performance when grouping fish populations in pelagic habitat (64%) rather than pooling all populations (36%) (Table 4.5). The formula for calculating the percentage is based on the best recruitment prediction achieved from different groups of populations (water column and all populations) obtained from the total test points found in the water column (pelagic or demersal). Similarly, I fit $\mathcal{M}_1$ ($\gamma < 0$) to demersal populations (i.e. 59 populations) where I found an improvement by grouping fish populations in demersal habitat (61%) rather than pooling all populations (39%) (Table 4.5). Results show that a more accurate recruitment predictions is obtained when limiting the analysis to species in the same water-column depth community.

4.5 Discussion

BHMs are found to reduce the uncertainty in key parameters and to provide a more accurate prediction of fish recruitment compared to non-BHMs (Table 4.1 and 4.2). The non-constant variance model ($\mathcal{M}_1$) showed to be more accurate than the other models (Table 4.4) and proved that the non-constant variance η_1 to be reliable over an approximate 95% credible interval in Faroe Plateau and

Table 4.5: Comparison of best recruitment prediction achieved by grouping populations by pelagic (open water), all populations, and demersal (bottom dwelling) habitats. The rows describe the grouping by water column (i.e. pelagic or demersal). Two experiments are conducted accordingly: first, predict recruitment by restricting the populations to the same water column; second, predict recruitment by pooling all populations. However, the vertical columns assigns the best prediction for each case respectively.

Water-column	Pelagic	All Populations (Pelagic + Demersal)	Demersal
Pelagic (total of 184 test points)	64%	36%	–
Demersal (total of 392 test points)	–	39%	61%

70% in North-East Arctic areas (Figures 4.6 (a) and (b)).

In this research I applied four different BHMs onto five different geographical regions (Celtic Sea, Faroe Plateau, Georges Bank, North-East Arctic and North Sea) and three other macro scale marine column zones (pelagic, demersal and all populations) so as to find the model that best describes the recruitment variability in these regions. The macro scale marine analysis is aimed to assess the influence of grouping fish species, across different water depths, on the prediction accuracy of fish recruitment. This research also extends the work of (Panikian et al., 2015) by applying a BHM for assessing the sign of the coefficient of heteroscedasticity η_1.

To validate this work, I compared the inferred S-R model parameters via BHM and non-BHM methods applied on the North-East Arctic area, which contains five stocks, where I found an increase of accuracy (or low uncertainty) in approximating the 95% credible interval using the BHM method. Additionally, I compared the prediction accuracy of fish recruitment using these two methods applied on the same area where I found a more accurate result is achieved via a BHM method. I repeated the same experiment on the Georges Bank (containing six populations) area and I observed a similar conclusion. This tells us that BHM are more accurate (in both inference and prediction) than non-BHM (or single species assessment) and worth considering for fairly short and noisy interdependent ecological time series data.

From a different perspective, results showed that the coefficient of heteroscedasticity deserves consideration in Faroe Plateau, North-East Arctic, pelagic and demersal communities because it had a better predictive accuracy on the test set (Table 4.4). This is in contrast to the constant variance model ($\eta_1 = 0$) that

is found to have some significance for the Celtic Sea and the North Sea as it provided a higher prediction accuracy. The reliability of the sign of η_1 in the three regions (Faroe plateau, Georges Bank and North-East Arctic) is found to be consistent for an approximate credible interval of 95%, 12% and 70% respectively. I concluded that the recruitment variability in the Georges Bank area is not possible to be explained with a heteroscedastic model because of the large amount of uncertainty of η_1 presented in the NEFSC-YELLGB-1935-2008 population (Figure 4.6).

The analysis is based on proposing four different Bayesian hierarchical models characterised by different S-R relationship (section 4.2.2), but with the same set of prior and hyperprior distributions. These models are assessed by different means (DIC, MLL, and predictive approach) where I found the DIC to be more reliable than both MLL and the predictive approach as the evaluation process (on the validation set) showed that it predicted the sensible model within three out of eight cases of the test set. Because the Deriso-Schnute model presents a singularity at $\gamma = 0$, I partitioned the search space into three disconnected zones ($\gamma < 0, \gamma \approx 0, \gamma > 0$) so as to overcome this limitation. The four models: $\mathcal{M}_1$, $\mathcal{M}_2$, $\mathcal{M}_3$ and $\mathcal{M}_4$ are analysed along with the different constraints on γ. The assessment on the test set helped us to identify the models with best prediction values such as: $\mathcal{M}_1$ (with $\gamma > 0$ and $\gamma \approx 0$) is found to provide the best recruitment prediction for the Faroe Plateau, North-East Arctic, pelagic and demersal water columns; $\mathcal{M}_2$ (with $\gamma < 0$) provided best accuracy for the Georges Bank; $\mathcal{M}_3$ (with $\gamma < 0$) found to provide the best prediction for the Celtic sea; and the model $\mathcal{M}_4$ (with $\gamma > 0$) found to provide the best prediction for the North sea and all populations. I checked whether the prediction accuracy of fish recruitment is affected by pooling multiple populations across the water column zones. Results showed that estimating fish recruitment by restricting the analysis to species within their own community (i.e. pelagic or demersal) has a higher accuracy than pooling all fish populations together (Table 4.5).

Marine fish populations are difficult to manage because of complexities resulting from: imperfect understanding of the ecosystem, imperfect understanding of the marine biology, imperfect sampling of fish populations and imperfect mathematical modelling. However, the management of fish populations is improved regularly by monitoring the performance of stock assessment in relation to the target objectives and providing feedback so as to improve the system.

In this work I provide to fishery managers (or policy makers) a new way of assessing the size of fish recruitment where I advocate the importance of collecting as many fish stock assessments from the same community (specified by the same geographical region and water column) as they can, then employ the different models $\mathcal{M}_1$, $\mathcal{M}_2$, $\mathcal{M}_3$ and $\mathcal{M}_4$ that I proposed to find the one that best describes the S-R variability, which is assessed by the lowest RMSE prediction value. A more compelling reason for conducting stock assessment selection based on water column is because I found an increase in accuracy by grouping stocks by water column rather than pooling all populations together (Table 4.5).

BHM proved to be a successful approach in managing marine communities indicating that a community based structure can explain better the recruitment variability of fish populations than analysing a single stock assessment. On the ecosystem level I conclude that the sea surface is to some extent disconnected from the sea bed in the sense that the water column mixing does not affect enormously nutrient supplies of these two habitats: as I found that fish stock assessments are best analysed on their own community.

As future work, one can probably look to overcome the limitations of the Deriso-Schnute model and find ways to interpolate between the Beverton-Holt, Ricker and Schaefer models without falling to the singularity at $\gamma = 0$.

Chapter 5

End-To-End Statistical Modelling for Marine Ecosystems via Machine Learning

I provide a simple modelling framework, exploiting algorithms from dynamic Bayesian networks, for coupling environmental, planktonic and fisheries data to arrive at predictive ecosystem-scale models. This study is concentrated on ICES Division VIa, with monthly and annual observations spanning years 1960 to 2014. I propose two ecological models to represent the structural dependencies within the data: the first is designated to analyse the full data set and the second to analyse the planktonic and abiotic variables only. Both models have shown a better prediction on the unseen data when compared to six other off-the-shelf algorithms. Although the proposed autoregressive models are too simple to capture complex long-term interdependencies, their simplicity can reveal clear short-term relationships. In this work, I show that the data support single-species rather than community-based fisheries management. Moreover, by applying perturbations to simulate climate change, the models consistently predict large disruptions to the dinoflagellate and fish larvae communities in response to changes in temperature and salinity.
The objectives of this Chapter are the following:

- To assess whether modelling fish populations within a community context is more appropriate than conventional species-based fisheries management methods.

- To develop methods for understanding the impact of environmental changes on marine species.

5.1 Introduction

There has been a rapid rise in the development of system scale marine models over the past decade (Travers et al., 2007; Rose et al., 2010) incorporating dynamics from ocean physics to top trophic level of fisheries. These so-called end-to-end models typically combine sub-models from hydrodynamics (representing coastal water dynamics), bio-geochemistry (representing the utilisation of nutrients by the lower trophic level organisms) and fisheries (representing the higher trophic level organisms) (Rose et al., 2010).
Keyl and Wolff (2008) described the impacts of fishery pressure and environmental variability on fish populations and pointed out that the performance of fished stocks (i.e. survival, growth and reproduction) can better be described if environmental or climatic variability is incorporated into the fisheries models. However, end-to-end models require advances in data collection, modelling theory and computing power. For example, the time series observations of ocean properties are collected by remote sensing from satellites, whereas assessments for the Continuous Plankton Recorder (CPR) data and fish populations are being conducted through marine ecological surveys. Since there are no rules defining the required level of detail that should be considered in an end-to-end research, the choice of processes and organisms remains an open problem (Rose, 2012).
There are currently several established modeling frameworks. Travers et al. (2009) coupled the ROMS-$N_2P_2Z_2D_2$ and the OSMOSE models into a single framework to capture a new range of two-way dynamics between low trophic level and high trophic level models by simulating both top-down and bottom-up effects. Fennel (2010) developed a two-way interaction NPZDF model that couples the mass fluxes between the state variables of the lower and upper food web —dominated by two prey species (sprat and herring) and one predator (cod). The Atlantis modeling framework simulates ocean physics, nutrient cycling, ecology, and fishery dynamics in a single modelling framework (Fulton, 2001; Link et al., 2010).
There has been much interest in forging formal and quantitative linkages between climate and living marine resources (Hollowed et al., 2011; Stock et al., 2011) so as to understand the impacts of climate variability on fish and fisheries.

The methods range from single-species stock assessments to techniques for evaluating impacts at the multispecies or ecosystem level. Fromentin and Planque (1996) were the first to prove that the North Atlantic Oscillation (NAO) index is correlated with the abundance of two major zooplankton copepod species in the eastern Northeast Atlantic and the North Sea through both temperature and wind speed (i.e. turbulent mixing affecting the development rate of zooplankton); however this correlation apparently broke down after 1996. Marine ecosystems reveal complex responses to climate changes because the response at the higher trophic level is not always proportional to the magnitude of changes of environmental conditions (Ito et al., 2010).

End-to-end models are becoming steadily more dependent on the mechanistic representations of complex dynamic systems that often require calibrating many parameters. This leads to potential difficulties with mathematical stability and tractability due to increased model complexity (Travers et al., 2007; Wong, 2014). Trifonova et al. (2015) compared Bayesian network modelling approaches with latent variables to reveal species dynamics for 7 geographically and temporally varied areas within the North Sea. They also applied structure learning techniques to identify functional relationships such as prey-predator between trophic groups of species that vary across space and time. In a more recent work, Trifonova et al. (2017) used a dynamic Bayesian network model with a hidden variable and spatial autocorrelation to explore future productivity of different fish and zooplankton species in response to changes in temperature within the North Sea.

In this Chapter, I analyse the influence of environmental changes on planktonic and fish populations using multivariate autoregressive models so as to determine the evolution of the system over time. The advantage of the proposed revised ecological model (REMO) lies in its probabilistic framework, its capability for dealing with multivariate time-series data streams contaminated by noisy and incomplete samples, and its flexibility in incorporating new potential sources of data (e.g. new species and environmental factors). REMO can also deal with previously unobserved conditions and respond to possible impacts of climate change on the ecosystem. I demonstrate REMO on a data set composed of 18 random variables collected from three different sources: the National Oceanic and Atmospheric Administration (NOAA) from which we obtained the abiotic components, the Sir Alister Hardy Foundation for Ocean Science (SAHFOS) institute from which I obtained the biotic components and the International Council for

Table 5.1: Five ICES data sets found in the Northwest Coast of Scotland and Northern Ireland (Division VIa) taken from the assessment year 2015.

id	Fish Stock	Stock Description	Species	Period	Type
1.	cod-scow	Cod in Division VIa (West of Scotland)	*Gadus morhua*	1981-2014	Demersal
2.	had-346a	Haddock in Subarea IV and Divisions IIIa West and VIa (North Sea, Skagerrak and West of Scotland)	*Melanogrammus aeglefinus*	1972-2014	Demersal
3.	her-67bc	Herring in Divisions VIa and VIIb,c (West of Scotland, West of Ireland)	*Clupea harengus*	1960-2014	Pelagic
4.	meg-4a6a	Megrim in Divisions IVa and VIa	*Lepidorhombus*	1985-2014	Demersal
5.	whg-scow	Whiting in Division VIa (West of Scotland)	*Merlangius merlangus*	1981-2014	Demersal

the Exploration of the Sea (ICES) from which I obtained the fish populations. The selected geographical location is situated in the Northwest Coast of Scotland and Northern Ireland (ICES Division VIa) and bounded in an area defined by the following coordinates: (60°30' north latitude, 4°00' west longitude), (60°30' north, 5°00' west), (60°00' north, 4°00' west), (60°00' north, 12°00' west), (54°30' north, 12°00' west), (54°30' north, 4°00' west) and then back to the point of beginning (60°30' north latitude, 4°00' west longitude).

Two application-driven outcomes of this study are: (1) to assess whether modelling fish populations within a community context is more appropriate than conventional species-based fisheries management methods, and (2) to develop methods for understanding the impact of environmental changes on marine species.

5.2 Materials and methods

In this research I select five fish stock populations from the International Council for the Exploration of the Sea (ICES) database (`http://standardgraphs.ices.dk/stockList.aspx`), living in the Northwest Coast of Scotland and North Ireland (ICES Division VIa). The length and characteristics of the data on the fish population are summarised in Table 5.1. To understand the influence of abiotic and biotic factors on the abundance of fish populations, I consider an additional 13 variables (environmental and biological) along with fish populations, which can be described as follows:

1. Arctic Oscillation (AO) is a climate index describing the atmospheric circulation over the Arctic. Positive values of AO indicates that the polar vortex (i.e. the polar pressure circulation) is stronger compared to the air pressure at mid-latitudes. This creates a strong air flow circulation and confines the coldest air to the high latitudes. However, under negative AO conditions, the air travels from the North pole toward low pressure areas and brings cold weather to North America, Europe, and Asia.

2. The North Atlantic Oscillation (NAO) index is viewed as a dominant cause of climate change in the North Atlantic region. The National Oceanic and Atmospheric Administration (NOAA) states that strong positive phases of NAO tend to be associated with above-normal temperatures across northern Europe and are also associated with above-normal precipitation over northern Europe and Scandinavia. However, opposite patterns of temperature and precipitation anomalies are typically observed during strong negative phases of the NAO.

3. Sea surface temperature (SST) has a direct influence on phytoplankton growth and metabolic rates.

4. Wind speed (Wind) often increases with a high NAO index. Strong winds induce greater turbulence and lead to more intense vertical mixing which is important in controlling spring phytoplankton bloom (Sverdrup, 1953).

5. Salinity (SAL) is a significant factor in the growth and reproduction of fish populations (Lowe et al., 2012).

6. The Southern Oscillation index (SOI) gives an indication of the development and intensity of El Niño or La Niña events in the Pacific Ocean, which is computed using monthly mean sea level pressure anomalies at Tahiti and Darwin.

7. Fish larvae (FishLarvae) are the part of the zooplankton that eat smaller plankton. They are usually consumed by larger animals. The newly hatched fish larvae have a total length of around three millimeters and can swim poorly compared to the older larvae which swim faster. The larva period is relatively short, typically several weeks, during which the larva grows and changes its structure and form to become a juvenile fish.

8. Krill (Krill) are small crustaceans which feed on phytoplankton and zooplankton; they exhibit large daily vertical migrations, moving near the surface at night and in deeper waters during the day. This vertical migration (across the water column) for feeding and reproductive purposes make them extremely vulnerable to predators.

9. Copepods (Cope) are important species for global ecology and the carbon cycle because they are a major food source for small fish and other crustaceans such as krill.

10. Large Copepods (LargeCope) are likely to consume larger and more prey than small copepods.

11. Fish Eggs (FishEgg) are often released in the sunlit zone of the water column, usually less than 200 meters below the surface. Fish eggs cannot swim at all but rely on the large yolk sac they carry within the egg for nourishment.

12. Dinoflagellates (Dinoflage) are typically small-sized stress-tolerant species of plankton.

13. Diatoms (Diatom) are a major group of algae, and are among the most common types of phytoplankton which dominate the water surface during the spring bloom.

The gathered ecological data for the biotic and abiotic factors cover a period starting from 1960 till 2014; but do not preserve consistency of time interval for the five fish populations (Table 5.1). A strong variability of NAO is often associated either with an index value greater than 1.0 or less than -1.0. This strong variability contributes to changes in the atmospheric pressure, wind speed, precipitation, air temperature and sea surface temperature (SST) anomalies (Hurrell, 1995). I proceed up the trophic layers, beginning with phytoplankton, zooplankton, through to fish. Around the UK, primary production follows a seasonal pattern. Winter storms and strong winds vertically mix the water column, transporting nutrient concentrations to the surface. In spring, when the sunlight intensity increases, conditions are suitable to stimulate growth of the phytoplankton and allow photosynthesis —a phenomenon known as the spring bloom. Sverdrup (1953) showed that there must be a critical depth on the surface layer such that

blooming can occur only if the depth of the mixed layer is less than a critical value. This is known as Sverdrup's critical depth hypothesis. The spring bloom is generally dominated by diatoms, photosynthetic algae which can be used as indicators for monitoring the environmental aquatic conditions and water quality (Hering et al., 2006). In summer, as the surface layers warm, a thermocline can develop preventing vertical mixing and re-suspension of nutrients, the surface waters become nutrient-replete and nanoflagellates (competitor organisms) become dominant; however, in autumn, during periods with low nutrient concentrations, dinoflagellates (stress tolerators) species become dominant (Hansen et al., 1996). Often around the UK, autumn storms can cause mixing and break down stratification, allowing the re-suspension of nutrients. In autumn months, light levels are often sufficient to 'kick-start' a secondary late bloom of diatoms. Moving into winter, the phytoplankton production depletes as daylight hours become shorter. Irigoien et al. (2000) suggest a possible link between NAO index and phytoplankton species because of significant correlation found between the winter NAO (December to March) and the diatom abundance (April to May). Edwards et al. (2001) concluded that variability in phytoplankton biomass is affected by the sea surface temperature influenced by the North Atlantic Oscillation index. Zooplankton are generally tiny animals (the majority of which are Crustacea) and, as all plankton, they are weak swimmers and usually drift along with the currents. They can be classified by size or stage of development. Among the large number of zooplankton, I selected Fish Eggs, Fish Larvae, Large Copepods, Small Copepods and Euphausiids (or krill) datasets from the CPR survey database (Johns, 2015).

I chose three different methods to conduct this research: the first is to analyse the data through a yearly index; the second through both monthly and yearly indices (monthly + yearly); and the third through a monthly index.

For the yearly index analysis, I encountered missing values in both planktonic (presented in monthly samples) and fish species (presented in yearly samples). To get around this problem, I computed the annual mean value for the observed planktonic data by throwing away samples containing missing values (in case they exist). For example, if in a particular year there are two months of missing values, I omitted those months and averaged the abundance of the species based on the observed values only. In statistics, a listwise deletion involves omitting an entire record from analysis if any single value is missing. I used this approach

to remove incomplete assessments of fish populations such that after combining the entire data set (abiotic + biotic + fish populations) across the year index, we obtain a rectangular data set with no missing values. I also called this method *the yearly data structure*.
For the monthly index analysis, I used a linear interpolation method to impute the missing monthly samples in the planktonic data set, and a linear interpolation for the fish species taken between two yearly assessment samples so as to impute a value for each month respectively —I also called this method *the monthly data structure*.
Lastly, for the (monthly + yearly) index analysis, I used a linear interpolation to recover monthly missing values for the planktonic data and restricted the analysis on observed assessments for fish populations from which I omitted missing values of fish populations using a listwise deletion. This approach resulted in constructing a high dimensional dataset (33 samples $\times$ 168 dimensions).
Because the listwise deletion approach removed too many observed samples (33 out of 55 years), I decided in a second part of this research not to include fish populations and limit the analysis only to biotic and abiotic variables.

5.2.1 The Data

Data sets (1-2) were downloaded from the National Weather Service publicly available at (`http://www.cpc.ncep.noaa.gov/products/precip/CWlink/daily_ao_index/ao_index.html`) and (`http://www.cpc.ncep.noaa.gov/products/precip/CWlink/pna/nao.shtml`) respectively; data sets (3-5) are taken from the interactive time series explorer (NOAA, 2015) website, which allows the user to extract the data by specifying a set of geographical coordinates; data set (6) is downloaded from the publicly available web site: (`http://www.cpc.ncep.noaa.gov/data/indices/`); and data sets (7-13) are taken from (Johns, 2015). The data set covers a period ranging from 1960 till 2014.
In this work, the data were divided into three parts: training, validation and test sets with proportions of 60%, 20% and 20% respectively.

5.2.2 Multivariate autoregressive models

In this research, I applied different multivariate first order autoregressive algorithms (Least Angle Regression (LARS), G1DBN, Simone, Gaussian Process,

GeneNet and a Baseline method) to come up with a (better) revised ecological model (REMO) to explain the data, while enforcing sparseness in network connectivity. Amongst the first order autoregressive models, I proposed a Baseline method as a very basic autoregressive model (with a coefficient parameter equal to one and with a zero noise) that consists of predicting the next step ahead value to be equal to the present observation. A detailed description of these methods is provided in Chapter 2, section 2.6. As an evaluation metric, I used the root mean square error (RMSE) to measure the spread of predictions around the ground truth. There are also other possible metrics that one could use such as the mean absolute error.

5.2.3 Selecting tuning parameters for the models

Below is the method that I followed to tune the parameters for the different models:

LARS: The shrinking of the coefficient estimates towards zero is performed with Lasso (Tibshirani, 1994) whereas the shrinkage parameter is estimated using a cross validation method.

G1DBN: the threshold α_1 is chosen such that after the first step of inference all random variables are expected to have no more than one parent; however, the choice of α_2 is less problematic and can be chosen so as to lower the number of edges. I applied a heuristic approach to set α_1, I carried out inference with different values of α_1, such that $\alpha_1 = \{0.08, 0.1, 0.2, 0.3\}$. The best result is found for $\alpha_1 = 0.08$ which resulted in a higher accuracy. On the other hand, I chose a low α_2 threshold (α_2=0.09) so as to be confident in the selected edges.

Simone: to determine the best number of edges, I applied a heuristic approach by measuring the RMSE associated to different value of edges: 30, 50, 100 and 150 edges. The best network is found with a penalty corresponding to at most 30 edges.

Gaussian Process: I used the Radial Basis kernel function for the covariance matrix because it fits the data better than: Polynomial, Linear, Laplacian and Bessel kernels.

GeneNet: to determine the best number of edges, I applied a heuristic approach by measuring the RMSE associated to different value of edges: 30, 50, 100 and 150 edges. The best network is found with a penalty corresponding to at most 150 edges.

5.2.4 Revised Ecological Model (REMO)

REMO is a heuristic search algorithm used for learning the ecological network structure using a multivariate autoregressive model, which is described as in Algorithm 6. After building REMO on the validation set, I tested all models on the test set and evaluate them according to the RMSE metric.

Algorithm 6 REMO local search algorithm.

1: Split the data set into: training, validation and test sets.
2: Select six (more or less) off-the-shelf models: Least Angle Regression (LARS), G1DBN, Simone, Gaussian Process, GeneNet and a Baseline method.
3: Train all these models on the training set.
4: Test the models on the validation set.
5: **for** each variable j $\leftarrow$ 1:18 (i.e. NAO, AO, Diatom, ...) **do**
6: Find the model that produced the best one-step-ahead prediction on the validation set (evaluated with the RMSE metric).
7: Construct the coefficient of the autoregressive matrix $\mathbf{A}$ from which I inferred the interaction between variables. For example, if we assume a first-order autoregressive process $X(t) = \mathbf{A}.X(t-1)$, a non-zero element a_{ij} from $\mathbf{A}$ implies an arc from X_j to X_i.
8: Extend the obtained solution $\sigma_j \leftarrow \{a_{ij}\}$ to encompass non-linear interactions, using 'bilinear terms'.
9: $\delta_j \leftarrow$ Greedy-Local-Search(σ_j, RMSE, $\mathcal{O}$) where $\mathcal{O}$ = {delete, reverse, add} edges. This local search is described in Chapter 2, Algorithm 1.
10: **end for**
11: REMO $\leftarrow \{\delta_j\} \quad \forall j$
12: **return** REMO

5.2.5 Revised Ecological Model with listwise deletion (REMO1)

Models (e.g. linear and non-linear) are trained using listwise deletion method (remove the cases with missing values) so as to restrict the data to the assessed (or observed) fish populations, biotic and abiotic values, then are validated on the validation set from which we choose between models that best fit the data. The consequence of listwise deletion results in discarding much information and

restricting the data set to range from 1985 till 2014. The process of inferring the network structure of REMO1, for the yearly index, is described in Algorithm 6. The method consists of: (1) rectangularising with listwise deletion all collected variables; (2) picking the best fit model (LARS, G1DBN, Simone, Gaussian process, GeneNet and Baseline) applied onto the validation data set for each variable respectively; (3) extending the solution to include nonlinear interactions (in some cases) using 'bilinear terms' (Buchel and Friston, 1997) such that to model a hypothesized interaction between variables X_{SAL} and X_{SST} I formed a new bilinear variable $I = X_{SST} \times X_{SAL}$; and (4) applying a greedy local search algorithm, as described in Algorithm 1 (Chapter 2), which adjust the initial structure by randomly adding, reversing or deleting one edge at a time so as to improve the accuracy upon the validation set (i.e. to attain a lower RMSE). The desired output is a network structure (from the set of all possible structures) that best fits the validation set.
For example, REMO1 depicted that the krill at time $t+1$ is influenced by several other variables at time t, such that

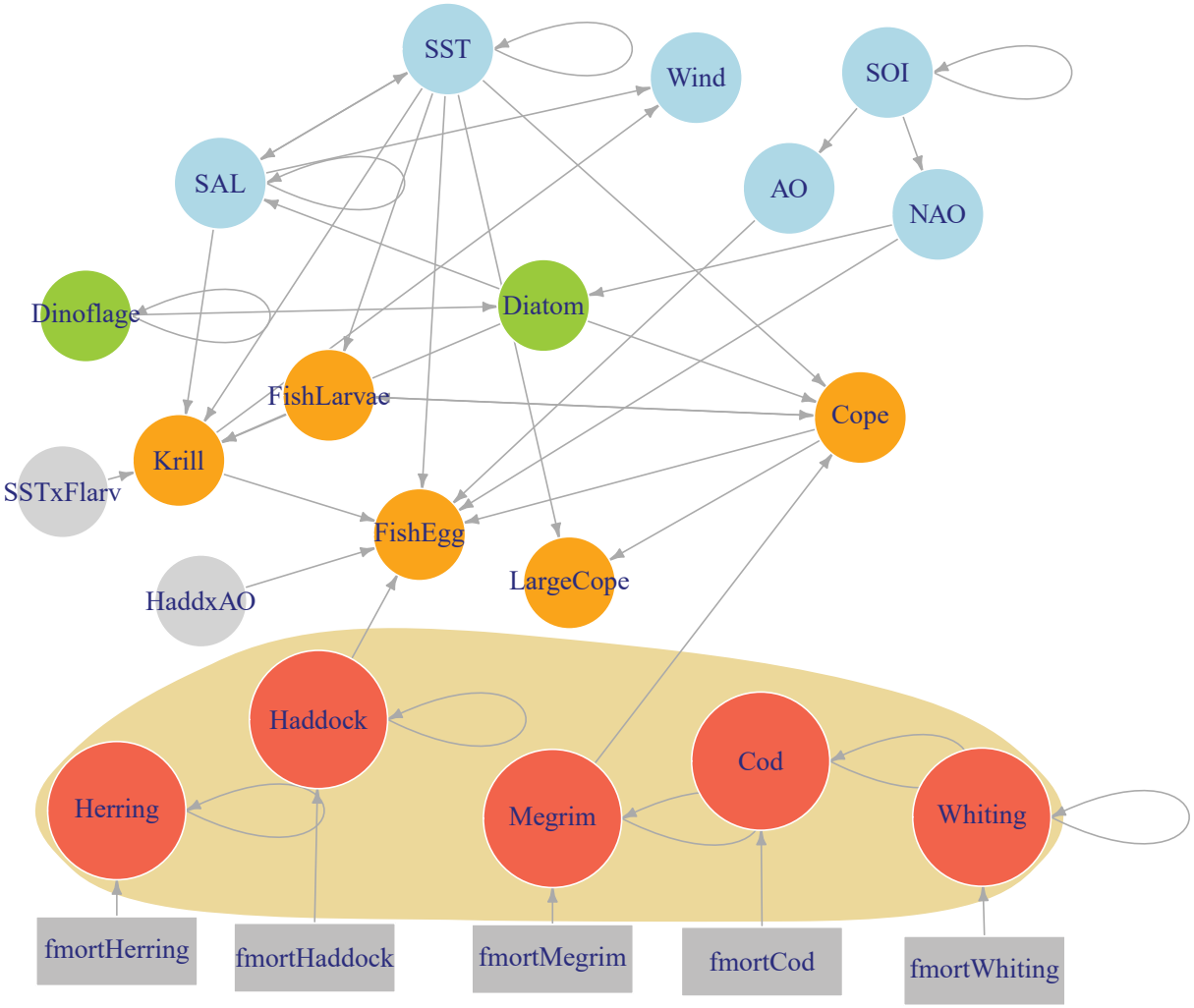
$$\begin{aligned} X_{\text{Krill}}(t+1) = \alpha_{13} X_{\text{SST}}(t) + \alpha_{14} X_{\text{Diatom}}(t) + \alpha_{15} X_{\text{FishLarvae}}(t) + \\ \alpha_{16} X_{\text{SAL}}(t) + \alpha_{17} X_{\text{SST}}(t) X_{\text{FishLarvae}}(t), \end{aligned} \tag{5.1}$$

where the α_i's are constant parameters fitted while training the model on the training data. The set of interactions that were discovered is described in Appendix F.1; however, the graphical representation of REMO1 is illustrated as in Figure 5.1. REMO1 manifests some unexpected relationships, including links from SAL and Krill to Wind, and Diatom to SAL. Obviously these links do not make sense from an ecological point of view but are picked up in the process of generating the model, as they provide a better fit to the data. I think that this issue is mainly caused by the fact that the annual mean data set with listwise deletion is small and noisy.

5.2.6 Bayesian hierarchical modelling for fish populations

In this section, I used previous models $\mathcal{M}_1$, $\mathcal{M}_2$, $\mathcal{M}_3$ and $\mathcal{M}_4$ described in chapter 4 on the fish populations and compared the predictive accuracy of fish recruitment on the test set. Results shows (Table 5.5) that the model $\mathcal{M}_2$ with $\gamma < 0$ outperforms the other models meaning that a Bayesian hierarchical model with

Figure 5.1: Graphical representation of REMO1 analysing the entire data set with listwise deletion for fish populations —ranging from 1985 to 2014. The nodes describe the random variables representing the ecological system; blue denotes abiotic variables, green denotes phytoplakton, grey denotes auxiliary variables (introduced for a better fit to the data), orange denotes zooplankton, red denotes fish populations, and finally the squares denote the fishing mortality rate. For a better fit, we introduced two auxiliary variables: SST × FLarvae and Haddock × AO. The edges with arrows describe dependencies among these variables. For example, an arrow from node SOI to node AO, describes a first order autoregressive model such that: $AO(t+1) = \alpha SOI(t)$.



community factor and heteroscedastic parameter provide the best accuracy for the ICES division VIa.

5.2.7 Revised Ecological Model applied to Biotic and Abiotic variables only (REMO2)

In this section, I decided to make use of all biotic and abiotic variables, without considering the five fish populations, so as to overcome limitations induced by listwise deletion. Another compelling reason for dropping the fish populations in this section is because I found previously that variations in climate have little influence on the stock size of fish populations. For example, REMO2 depicted that the krill at time $t+1$ is influenced by several other variables at time t, such that

$$X_{\text{Krill}}(t+1) = \alpha_{15} X_{\text{SST}}(t) + \alpha_{16} X_{\text{Cope}}(t) + \alpha_{17} X_{\text{FishLarvae}}(t).$$

The set of interactions that were discovered in this section is described in Appendix F.2; the graphical representation of the learned model is illustrated in Figure 5.2. I found that REMO2 did not develop unexpected relationships as in REMO1 because in this case the data set is almost two times bigger than that of REMO1.

5.2.8 Modelling perturbations to the Ecological system

Climate change is the main cause of increasing the sea surface temperatures and melting of glaciers on land (in places like Antarctica and Greenland). To predict the possible consequence of these adverse situations on the ecological system, I simulated future responses from REMO1 and REMO2 according to a set of scenarios described as follows:

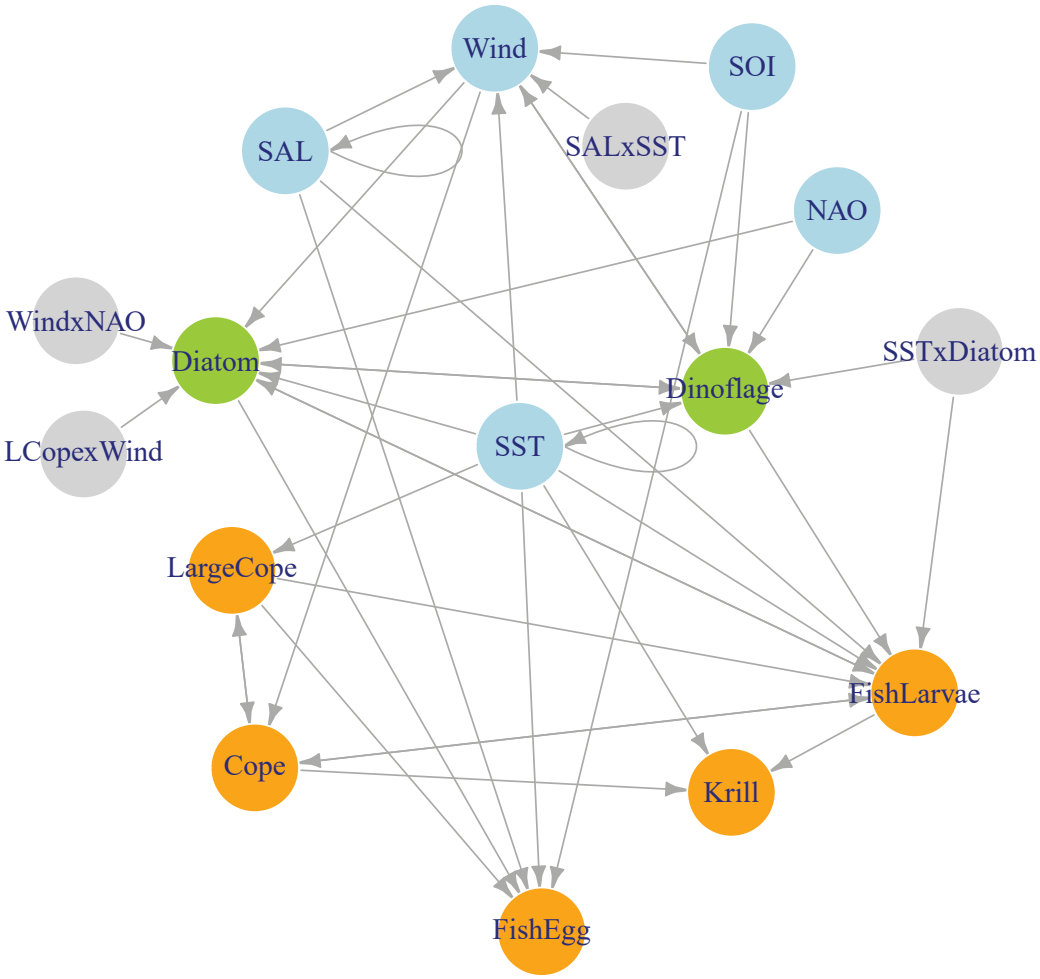
S1: annual mean SST increases by 2°C.

S2: annual mean Salinity decreases by 5%.

S3: annual mean SST increases by 2°C and annual mean Salinity decreases by 5%.

S4: annual mean NAO hits an index of +1.

Figure 5.2: Graphical representation of REMO2 with biotic and abiotic variables only —ranging from 1960 to 2014. The nodes describe the random variables representing the ecological system; blue denotes abiotic variables, green denotes phytoplakton, grey denotes auxiliary variables (introduced for a better fit to the data), and orange denotes zooplankton. For a better fit, we introduced four different auxiliary variables: Wind × NAO, LCope × Wind, SAL × SST and SST × Diatom. The edges with arrows describe dependencies among these variables. For example, the self-loop on the SAL node describes a first order autoregressive model such that: $SAL(t+1) = \alpha SAL(t)$.



S5: annual mean NAO falls to an index of -1.

S6: annual mean Wind speed reaches 10.8 m/s (an increase of 20% from 2014's record).

S7: annual mean SOI hits an index of +1.

S8: annual mean SOI falls to an index of -1.

5.3 Results

In accordance with the above plan, results are split in three sections: the first is for REMO1, the second is for the Bayesian hierarchical modelling to assess the community influence on fish populations and the third is for REMO2.

5.3.1 Evaluating models with listwise deletion for fish populations (REMO1)

In this section I compare results generated by all models (i.e. LARS, G1DBN, Simone, Gaussian Process, GeneNet, Baseline and REMO) based on the listwise deletion data set. Firstly, I used the annual mean average of the data to compare the prediction accuracy for each model (Table 5.2): the smaller RMSE value means the higher the prediction accuracy. All models (i.e. LARS, G1DBN, Simone, Gaussian Process and GeneNet) except the Baseline model have their input parameters tuned as described in section 5.2.3.
REMO1 predicted with a higher level of accuracy on the validation set (15 out of 18 random variables) than the other models (Table 5.2), and it preserved its accuracy by showing a better prediction over the test set (Table 5.3). I interrogated REMO1 by measuring the percentage of change for each random variable when simulating its response to a set of scenarios (S1-to-S8) so as to understand the possible impact of abiotic variables on the ecological system. The percentage change is described as $\{(Si - S0)/S0\} \times 100$ so as to produce the result in percent (%), $E0$ is the prediction under normal condition and Si is the prediction under conditions (S1-to-S8). There are some instances where REMO1 and REMO2 predicted negative numbers indicating extinction of species after perturbation, this interpretation is caused by the fact that the models are linear and do not

Table 5.2: Root Mean Square Error (RMSE) characteristic of models —LARS, G1DBN ($\alpha_1 = 0.08$), Simone (with 30 edges), Gaussian Process (RBF Kernel), GeneNet (150 edges), Baseline and REMO1—applied to the Validation set of the yearly data structure with listwise deletion. The highlighted cells represent the best RMSE values for each random variable respectively.

Variable	**L**ARS	**G**1DBN	**S**imone	**G**aussian Process	**G**eneNet	**B**aseLine	**R**EMO1
AO	0.241	0.239	0.241	**0.231**	0.239	0.273	0.234
NAO	0.295	0.275	0.268	0.293	0.281	0.351	**0.243**
SST	0.151	**0.119**	0.138	0.418	0.368	0.138	0.138
Wind	0.471	0.422	0.365	0.415	0.517	0.466	**0.322**
SAL	0.070	0.095	0.071	0.092	**0.054**	0.071	0.071
SOI	1.376	0.506	0.516	0.464	0.529	**0.393**	**0.393**
FishLarvae	0.172	0.168	0.118	0.138	0.238	0.199	**0.114**
Krill	0.608	0.600	0.534	0.687	0.551	0.673	**0.505**
LargeCope	5.504	6.193	6.851	6.484	8.332	10.327	**4.895**
FishEgg	0.191	0.193	0.159	0.195	0.161	0.244	**0.153**
Dinoflage	9930.857	9966.763	10270.354	11334.991	13182.095	**9589.969**	**9589.969**
Diatom	14277.515	14194.859	35777.813	15555.976	37000.384	35777.813	**11069.484**
Cope	216.743	222.094	132.510	272.516	193.485	199.741	**129.596**
Cod	3298.835	**2267.984**	**2267.984**	17499.467	4301.217	**2267.984**	**2267.984**
Haddock	639829.627	503027.757	**341845.100**	494502.263	359136.432	**341845.100**	**341845.100**
Herring	69603.857	**64365.717**	**64365.717**	78882.110	69765.521	**64365.717**	**64365.717**
Megrim	0.382	**0.111**	0.111	0.221	0.205	**0.111**	**0.111**
Whiting	13370.767	**2660.310**	**2660.310**	38500.380	5257.545	**2660.310**	**2660.310**

Table 5.3: Root Mean Square Error (RMSE) characteristic of models —LARS, G1DBN ($\alpha_1 = 0.08$), Simone (with 30 edges), Gaussian Process (RBF Kernel), GeneNet (150 edges), Baseline and REMO1—applied to the Test set of the yearly data structure with listwise deletion. The highlighted cells represent the best RMSE values for each random variable respectively.

Variable	**L**ARS	**G**1DBN	**S**imone	**G**aussian Process	**G**eneNet	**B**aseLine	**R**EMO1
AO	0.555	0.524	0.539	0.524	0.524	0.792	**0.519**
NAO	0.822	0.657	0.539	0.614	0.724	0.810	**0.536**
SST	**0.150**	0.157	0.180	0.383	0.308	0.180	0.180
Wind	0.508	0.335	0.255	**0.232**	0.303	0.415	0.293
SAL	0.143	0.101	0.083	0.121	0.092	**0.083**	**0.083**
SOI	0.899	0.687	**0.548**	0.645	0.687	0.791	0.791
FishLarvae	0.257	0.126	0.121	0.092	0.124	0.119	**0.085**
Krill	1.037	0.712	0.716	**0.633**	0.817	1.081	0.730
LargeCope	**9.388**	9.524	11.528	10.244	11.274	15.075	11.814
FishEgg	**0.198**	0.215	0.245	0.206	0.209	0.300	0.242
Dinoflage	13013.320	14055.920	12988.625	11569.258	11565.062	**8089.141**	**8089.141**
Diatom	30140.735	30843.110	47201.027	31617.540	**29092.234**	47201.027	32225.974
Cope	202.506	182.532	277.824	246.966	262.394	**87.147**	243.824
Cod	3779.347	**665.433**	**665.433**	18626.226	7393.712	**665.433**	**665.433**
Haddock	892128.708	332881.220	**208951.922**	670634.096	367830.331	**208951.922**	**208951.922**
Herring	96843.352	**72627.553**	**72627.553**	281908.596	177245.396	**72627.553**	**72627.553**
Megrim	0.540	**0.183**	**0.183**	0.407	0.491	**0.183**	**0.183**
Whiting	6526.765	**3364.011**	**3364.011**	33051.149	4366.370	**3364.011**	**3364.011**

Table 5.4: Percentage change of variables predicted by REMO1: applied to list-wise deletion for fish populations with yearly data structure ranging between 1985 and 2014. S0: one-step-ahead prediction in normal conditions based on 2014 values; S1: increased SST by 2°C; S2: decrease Salinity by 5%; S3: S1 and S2 observed simultaneously; S4: NAO +1; S5: NAO -1. The cells containing the value 'Extinct' designate variables that are predicted to become extinct after perturbation; in reality this is more likely to indicate community-scale reorganisation.

Variable	**S**0	**S**1(%)	**S**2(%)	**S**3(%)	**S**4(%)	**S**5(%)
SST	11.3283333	17.65	0	17.65	0	0
Wind	9.426231376	0	-5.09	-5.09	0	0
SAL	35.07677523	0	-5	-5	0	0
SOI	-0.133333333	0	0	0	0	0
FishLarvae	0.227383987	7.26	0	7.26	0	0
Krill	1.087443078	-89.53	-43.88	Extinct	0	0
LargeCope	23.19401715	149.38	0	149.38	0	0
FishEgg	0.158427557	23.60	0	23.60	Extinct	193.92
Dinoflage	26456.31641	0	0	0	0	0
Diatom	79622.74432	0	0	0	15.75	-22.87
Cope	602.1739492	-3.89	0	-3.89	0	0

incorporate knowledge to avoid negative predictions. Table 5.4 shows that under conditions (S1-to-S5) I observe a big impact in the planktonic community, whereas under conditions S3 and S4, we reach a concerning ecological level showing an extinction of Krill and fish eggs. I discarded conditions S6, S7 and S8 because they were found to have no effect on the ecological system modelled by REMO1; I also omitted fish populations from Table 5.4 because changes in environmental conditions are found to have no immediate impact on fish stocks, as justified by the multivariate first-order autoregressive model (Figure 5.1).

Secondly, I applied the monthly data structure for learning the models for which each row represents a monthly sample: abiotic and plankton data are sampled on a monthly basis, but fish stocks are linearly interpolated between two consecutive annual estimations so as to obtain an estimate for each month respectively. I found LARS outperformed the other models providing the best RMSE values over the majority of the variables; but when I interrogated the model the forecasts were not reliable because the model predicted an extinction of larvae, copepods and dinoflagellates under normal conditions (S0), which indicated it to be unreliable.

Thirdly, I applied the (monthly + yearly) data structure: I preserved the monthly frequency for abiotic and plankton data and the yearly frequency for the fish

Table 5.5: Predictive approach for evaluating models $\mathcal{M}_1$, $\mathcal{M}_2$, $\mathcal{M}_3$ and $\mathcal{M}_4$ applied to the testing set of: Herring, Haddock, Cod and Whiting populations. The numbers scored in a model represent the cases for which this particular model is found to minimise the RMSE.

	$\mathcal{M}_1$			$\mathcal{M}_2$			$\mathcal{M}_3$			$\mathcal{M}_4$		
Predictive approach	$\gamma<0$	$\gamma\approx 0$	$\gamma>0$	$\gamma<0$	$\gamma\approx 0$	$\gamma>0$	$\gamma<0$	$\gamma\approx 0$	$\gamma>0$	$\gamma<0$	$\gamma\approx 0$	$\gamma>0$
ICES VIa	5	0	3	**12**	8	10	8	0	5	7	6	11

species. This approach led to a high dimensional data (33 samples $\times$ 168 dimensions) with limited training set samples. The Baseline method showed a better accuracy over the validation set than the other models, but when evaluating it on the testing set it reported a poor accuracy (results are not displayed). Since the Baseline model cannot reflect an immediate change (at the next time step) in response to climate changes, I instead interrogated the LARS model's response to those changes. This showed it to be unreliable as it predicted extinction of some species under normal conditions (S0).

5.3.2 Bayesian hierarchical modelling applied on fish populations

In this section, I applied the different models ($\mathcal{M}_1$, $\mathcal{M}_2$, $\mathcal{M}_3$ and $\mathcal{M}_4$) described in Chapter 4 on fish stock and recruitment datasets. I excluded the Megrim population from this experiment because it does not contain recruitment assessment values. Table 5.5 describes the model $\mathcal{M}_2$ with $\eta_1 \neq 0$, $\gamma < 0$ and community factor as the best predictive accuracy model on fish recruitment.

5.3.3 Evaluating models with Biotic and Abiotic variables (REMO2)

In this section I applied the evaluation method only to the yearly data structure, provided that the other structures (i.e. (monthly + yearly) and monthly) are ruled out in the previous section (section 5.3.1). I trained all the models on the training set, then validated them on the validation set. I followed the same approach as before to pick the best fit model for the purpose of building REMO2, as illustrated in Figure 5.2. Table 5.6 shows that REMO2 outperformed the other models on the validation set by recording a best fit for 7 out of 13 variables; however, Table 5.7 shows a domination of REMO2 over the other models on the

Table 5.6: Root Mean Square Error (RMSE) characteristic of models —LARS, G1DBN ($\alpha_1 = 0.1$), Simone (with 100 edges), Gaussian Process (RBF Kernel), GeneNet (30 edges), Baseline and REMO2—applied to the Validation set of the yearly data structure with biotic and abiotic variables only. The highlighted cells represent the best RMSE values for each random variable respectively.

Variable	**L**ARS	**G**1DBN	**S**imone	**G**aussian Process	**G**eneNet	**B**aseLine	**R**EMO2
AO	0.282	0.269	0.288	**0.234**	0.269	0.368	0.255
NAO	0.289	0.290	0.288	**0.286**	0.315	0.413	0.288
SST	0.338	0.328	0.234	0.327	0.310	**0.234**	**0.234**
Wind	0.411	0.388	**0.305**	0.386	0.409	0.358	0.344
SAL	0.076	0.083	0.083	**0.069**	0.080	0.083	0.083
SOI	0.590	0.598	0.687	0.558	**0.550**	0.687	0.582
FishLarvae	0.290	0.240	0.262	0.274	0.275	0.318	**0.194**
Krill	0.849	0.828	0.954	0.846	0.796	1.121	**0.794**
LargeCope	**11.084**	11.186	14.287	11.391	13.263	14.287	11.608
FishEgg	0.198	0.192	0.203	0.194	0.191	0.203	**0.177**
Dinoflage	14939.393	12642.507	13183.215	15832.408	14351.051	13458.519	**11770.892**
Diatom	24083.318	35609.669	35609.669	24800.425	22158.042	35609.669	**21722.851**
Cope	259.386	350.382	350.382	298.497	323.584	350.382	**255.777**

Table 5.7: Root Mean Square Error (RMSE) characteristic of models —LARS, G1DBN ($\alpha_1 = 0.1$), Simone (with 100 edges), Gaussian Process (RBF Kernel), GeneNet (30 edges), Baseline and REMO2—applied to the Test set of the yearly data structure with biotic and abiotic variables only. The highlighted cells represent the best RMSE values for each random variable respectively.

Variable	**L**ARS	**G**1DBN	**S**imone	**G**aussian Process	**G**eneNet	**B**aseLine	**R**EMO2
AO	0.425	**0.398**	0.403	0.430	0.398	0.618	0.406
NAO	0.455	0.447	**0.424**	0.499	0.550	0.652	0.438
SST	0.485	0.396	**0.153**	0.479	0.375	**0.153**	**0.153**
Wind	0.523	0.448	**0.382**	0.484	0.526	0.450	0.433
SAL	0.094	0.073	0.073	0.098	**0.067**	0.073	0.073
SOI	0.591	0.591	0.649	0.605	**0.556**	0.649	0.639
FishLarvae	0.315	0.147	0.184	0.336	0.203	0.157	**0.136**
Krill	0.904	0.788	0.737	0.828	0.766	0.909	**0.685**
LargeCope	8.862	**8.510**	12.174	9.439	8.532	12.174	9.410
FishEgg	0.216	0.196	0.278	0.210	0.194	0.278	**0.163**
Dinoflage	23798.406	14351.058	11182.163	22548.823	12441.963	7542.022	**7483.482**
Diatom	36701.539	36824.063	36824.063	33403.342	33520.037	36824.063	**23705.011**
Cope	268.143	108.716	108.716	318.847	247.794	**108.716**	258.466

Table 5.8: Percentage change of variables predicted by REMO2: restricted on biotic and abiotic data sets with yearly data structure ranging between 1960 and 2014. S0: one-step-ahead prediction in normal conditions based on 2014 values; S1: increased SST by 2°C; S2: decrease Salinity by 5%; S3: S1 and S2 observed simultaneously; S4: NAO +1; S5: NAO -1; S6: Wind speed increases by 20%; S7: SOI +1; S8: SOI -1. The cells containing the value 'Extinct' designate variables that are predicted to become extinct after perturbation; in reality this is more likely to indicate community-scale reorganisation.

Variable	**S**0	**S**1(%)	**S**2(%)	**S**3(%)	**S**4(%)	**S**5(%)	**S**6(%)	**S**7(%)	**S**8(%)
SST	11.3283333	17.65	0	17.65	0	0	0	0	0
Wind	9.526358927	18.21	-36.95	-183.55	0	0	0	0.04	-0.03
SAL	35.07677523	0	-5	-5	0	0	0	0	0
FishLarvae	0.225534979	Extinct	Extinct	Extinct	0	0	0	0	0
Krill	1.250787427	8.36	0	8.36	0	0	0	0	0
LargeCope	18.52169331	68.49	0	68.49	0	0	0	0	0
FishEgg	0.255274549	149.75	37.27	187.02	0	0	0	-7.85	6.00
Dinoflage	17920.67957	Extinct	0	Extinct	36.65	-53.20	Extinct	44.11	-33.73
Diatom	90137.13516	23.70	0	23.70	5.58	-8.11	-9.18	0	0
Cope	559.1973576	0	0	0	0	0	-36.22	0	0

test set and showing a higher prediction accuracy for 6 out of 13 variables. I interrogated REMO2 by simulating its response on a set of scenarios (S1-to-S8). I did not include in this table the forecasts for AO, NAO and SOI because they are modelled as zero (Appendix F.2); I noticed that under conditions (S1-to-S8) there is a big impact throughout the ecosystem. In particular, under S3, I found a serious risk of extinction for both fish larvae and dinoflagellates species due to an increase of SST and a decrease in salinity, showing a bigger effect on the extinction of fish larvae found under S1. However, conditions (S4, S5 and S6) exert only an influence on phytoplankton communities.

5.4 Discussion

REMO1 depicted a single species fish population analysis rather than a community based interaction (Figure 5.1); in contrast to the Bayesian hierarchical models which depicted a community based structure (see section 5.2.6) as a more accurate model for predicting fish recruitment. REMO2 predicted a disappearance of fish larvae and loss of dinoflagellate blooms in response to an increase of sea surface temperature by 2°C (Table 5.8).

Over the last ten years, there have been a large number of different modelling approaches applied to end-to-end modelling (Rose et al., 2010; Fulton, 2010). My

objective was to provide a statistical model capable of analyzing marine ecosystem response to environmental perturbations. The model consists of analysing marine ecosystem data from several sources, with different temporal resolutions, bounded by the Northwest Coast of Scotland and Northern Ireland (Division VIa). This geographical area can have temperate oceanic plankton taxa as well as colder water plankton as the shelf-edge current can bring warmer water species around the top of Scotland and into the North Sea.

I collected 13 different random variables for both biotic and abiotic factors and five additional ICES fish populations from which I derived two sets of data: the first is based on a listwise deletion for the fish populations, and the second is based on analysing the biotic and abiotic variables only. Since there is no general theory for how to combine processes and organisms that operate on different time and space scales together into a model, especially when one goes up the trophic levels, I applied a set of first order autoregressive methods to infer the relationship among the random variables so as to understand the influential factors affecting each variable. I proposed a method for structural learning and for predicting one-step-ahead values for all the ecological variables, which showed a higher accuracy rate than the other statistical models (i.e. LARS, G1DBN, Simone, Gaussian Process, GeneNet and Baseline).

In addition to these two methods, I looked into imputing missing fish stock assessments using the bootstrap expectation maximisation (EMB) algorithm (Honaker and King, 2010). This imputation method builds on the concept of multiple imputation that consists of extracting relevant information from the observed portions of a data set so as to impute multiple values for each missing sample, but when I compared it against the listwise deletion approach I found the latter providing a higher accuracy over biotic and abiotic data, hence the extrapolation of missing fish stock assessments was omitted from this research.

I found the data set arranged on an annual mean basis (yearly data structure) produced a more comprehensible Markovian network structure than the (monthly + yearly) and monthly data structures. I then simulated the response from REMO with a set of different scenarios (S1-to-S8) so as to understand the impact of weather change on the marine ecosystem; the model flagged serious risk values for fish larvae, fish eggs, krill and dinoflagellates under the following cases: S1, S2, S3, S4 and S6 (Table 5.4 and 5.8). Figure 5.1 illustrates that fish stocks are independent of each other within this modelling context, supporting the analysis

of fish stocks (in isolation) without considering the effect of the community. This was one of the reasons why I omitted fish populations in the second part of this research where I repeated the analysis over the biotic and abiotic variables only. The advantage of this approach is that I used as much information as possible (ranging from 1960 to 2014).

Table 5.4 and 5.8 showed that the number of fish eggs appear to be increasing under the effect of temperatures. I explain this behaviour as the fecundity of female fish increases with water temperatures. The abundance of diatoms and large copepods are found to increase under warmer water temperatures; but the abundance of copepods is found to decrease with temperature (Table 5.8). This is consistent with the theory that *Calanus finmarchicus* and *Calanus helgolandicus* copepods would occupy a niche (and grow) in cold and warm water temperatures respectively (Bonnet et al., 2008); these two species can co-occur in the same geographical region such that the former species becomes more abundant in cooler temperatures earlier in the year and the latter species becomes more abundant in warmer temperatures later in the year. Table 5.8 shows that an increase of 20% in wind speed could destabilise the ecosystem by reducing the number of diatoms and copepods. However, the model did not show a consistent signal (Table 5.4 and 5.8) concerning the effect of temperature on Krill. The most serious risk remains that of the sea surface temperature where an increase by 2°C could jeopardize the fish larvae in the short run, and hence deplete the fish stock size on the long run.

This study shows that statistical analysis using multivariate first-order autoregressive models can be helpful in revealing the underlying relationship among the random variables so as to analyse the ecosystem response to environmental perturbations.

Both REMO1 and REMO2 sometimes predict negative values for variables in response to ecological perturbations. This limitation is caused by the linear regression models that has been employed, which might be overcome with some non-linear models. Moreover, I explain the prediction of a negative value by destabilizing the ecosystem equilibrium point (nutrient diffusion) where I can't say what will happen, but I think that the ecological system will have to reorganise itself in some way, involving zooplankton, phytoplankton and fish species. REMO is limited for predicting one-step-ahead forecasts of ecological variables

(Figure 5.1 and 5.2) and may be helpful for formulating end-to-end fisheries management. Possibly, one can extend this work to cover long term forecasts through a multivariate non-linear regression analysis.

Chapter 6

Conclusion and outlook

In this final chapter, I summarise the principal findings and contributions of my research, then I review each chapter of this thesis and finally I identify possible avenues for future work.

6.1 Thesis contribution

This thesis contributes to fishery management with three important inputs: (i) the non-constant variance parameter is important in analysing the stock-recruitment relationship in a single fish populations, provided the consistency in sign; (ii) Bayesian hierarchical models based on the Deriso-Schnute relationship is a sensible choice for forecasting fish recruitment; (iii) marine communities will become endangered if the sea surface temperature increases by 2°C, which could cause extinction of fish eggs and loss of dinoflagellate blooms.

6.2 Thesis summary

In Chapter 1, I defined the research problem and methods to be used for completing this project where I introduced the Deriso and Schnute model to lay down the foundation for modelling the growth of fish populations.
In Chapter 2, I reviewed basic probability theory, frequentist and Bayesian paradigms, dynamic Bayesian networks, non-parametric models and time series modelling for handling uncertainties in fish populations.
In Chapter 3, I developed and implemented methods for identifying the non-constant variance (heteroscedasticity) in the spawner-recruit relationship. I found

heteroscedastic models tend to fit the S-R model inputs better than constant variance models across the majority of stocks, and strong evidence for a negative coefficient of heteroscedasticity in seven cases (Table A.1), including exploited *cod*, *herring* and *whiting* stocks in addition to *olive flounder* and *Peruvian anchoveta*. I advocate that the non-constant variance parameter in these cases deserves to be taken into account by managers. In contrast, only one stock was identified as having a positive coefficient of heteroscedasticity at the 95% confidence levels.
To determine whether I can reliably estimate the sign of η_1, I tested whether the confidence interval lies in a region showing a consistent sign with the coefficient where I found that both frequentist and Bayesian methods led approximately to equivalent inference.
To reliably identify a negative coefficient of heteroscedasticity, managers or fisheries scientists using the frequentist methods should check that their chosen confidence interval lies in the negative region; those using the Bayesian framework can consider the proposed priors (i.e. π_1 or π_2) as a non-informative benchmark prior and check whether their Bayesian credible interval lies in the negative region. I note that Bayesian approaches may be particularly useful where priors can be specified based on information about similar stocks in other locations. To protect this work against false positives or negatives, I recommend fisheries scientists to use both frequentist and Bayesian methods when assessing stocks for heteroscedasticity; if both methods agree then there would be strong support for our conclusion being correct; otherwise they should investigate the limitation of each method separately.
In Chapter 4, I extended the analysis of Chapter 3 by applying a Bayesian hierarchical model for assessing the sign of the coefficient of heteroscedasticity η_1. I found that the Bayesian hierarchical model applied to fish stocks living in a community has reduced the uncertainty in parameter estimates of the S-R relationship compared to the case where the analysis is based on a single stock assessment. I proposed four different S-R relationships based on the Deriso-Schnute (Deriso, 1980; Schnute, 1985) model. These models ($\mathcal{M}_1$, $\mathcal{M}_2$, $\mathcal{M}_3$ and $\mathcal{M}_4$) are assessed on five different geographical areas: Celtic sea, Faroe Plateau, Georges Bank, North-East Arctic and North sea; and three other macro scale marine column zones: pelagic, demersal and all populations. The macro scale marine analysis is aimed to assess the influence of grouping fish species, across different water depths, on the prediction accuracy of fish populations.

Results showed that the coefficient of heteroscedasticity deserves consideration in Faroe Plateau, North-East Arctic, pelagic and demersal communities because it had a better predictive accuracy on the test set; in contrast to the constant variance model ($\eta_1 = 0$) that is found to have some significance for the Celtic Sea, the North Sea and all populations as it provided a higher prediction accuracy. The reliability of the sign of η_1 is found to be approximately consistent in two regions: Faroe Plateau and North-East Arctic with an approximate credible interval of 95% and 70% respectively.

Because the Deriso-Schnute model presents a singularity at $\gamma = 0$, I partitioned the search space into three disconnected zones ($\gamma < 0, \gamma \approx 0, \gamma > 0$) so as to overcome this limitation. The models are analysed along with different constraints on γ; whereas the assessment on the test set helped us to identify the model with best prediction values. The model $\mathcal{M}_1$ (with $\gamma > 0$ and $\gamma \approx 0$) is found to provide the best recruitment prediction for the Faroe Plateau, North-East Arctic, pelagic and demersal water columns; $\mathcal{M}_2$ (with $\gamma < 0$) provided best accuracy for the Georges Bank; $\mathcal{M}_3$ (with $\gamma < 0$) found to provide the best prediction for the Celtic sea; and the model $\mathcal{M}_4$ (with $\gamma > 0$) found to provide the best prediction for the North sea and all populations. I checked whether the prediction accuracy of fish recruitment is affected by pooling multiple populations across the water column zones; my results showed that estimating fish recruitment by restricting the analysis to species within their own community (i.e. pelagic or demersal) has a higher accuracy than pooling all fish populations together.

In this work, I provide to fishery managers (or policy makers) a new way of assessing the size of fish recruitment: I advocate the importance of collecting as many fish stock assessments within the same community as they can, then apply them to the four (or other) possible models so as to check the one that has the lowest RMSE value in recruitment prediction. A more compelling reason to conduct stock assessment selection based on water column is because I found an increase in accuracy compared to the case of pooling pelagic and demersal populations together.

On the ecosystem level I conclude that the sea surface is to some extent disconnected from the sea bed in the sense that the water column mixing does not affect enormously nutrient supplies of these two habitats: I found that fish stock assessments are best analysed on their own community.

In Chapter 5, I aimed to provide a statistical model capable of analysing marine ecosystem response to environmental perturbations. The model consisted of analysing marine ecosystem data from several sources, with different temporal resolutions, bounded by the Northwest Coast of Scotland and Northern Ireland (Division VIa). This geographical area can have temperate oceanic plankton taxa as well as colder water plankton as the shelf-edge current can bring warmer water species around the top of Scotland and into the North Sea.

I collected 13 different random variables for both biotic and abiotic factors and five additional ICES fish populations from which I derived two data sets: the first is based on a listwise deletion for the fish populations, and the second is based on analysing the biotic and abiotic variables only. Since there is no general theory on how to combine processes and organisms that operate on different time and space scales together into a model, especially when one goes up the trophic levels, I applied a set of first order autoregressive models from which I built a revised ecological model (REMO) that outperformed the other models. I used REMO to simulate future responses with a set of different enquiries (E1-to-E8) to understand the impact of weather change on the marine ecosystem; the model showed serious risk values for fish larvae, fish eggs, krill and dinoflagellates when the sea surface temperature increases by 2°C.

The analysis shows that the number of fish eggs appear to be increasing under the effect of temperatures, I explain this behaviour as the fecundity of female fish increases with water temperatures; the abundance of diatoms and large copepods are found to increase under warmer water temperatures; but the abundance of copepods are found to decrease with temperature. This is consistent with the theory that *Calanus finmarchicus* and *Calanus helgolandicus* copepods would occupy a niche (and grow) in cold and warm water temperatures respectively (Bonnet et al., 2008); these two species can co-occur in the same geographical region such that the former species becomes more abundant in cooler temperatures earlier in the year and the latter species becomes more abundant in warmer temperatures later in the year. The results showed that an increase of 20% in wind speed could destabilise the ecosystem by reducing the number of diatoms and copepods. However, the most serious risk remains that of the sea surface temperature where an increase by 2°C could jeopardize the fish larvae in the short run; and hence deplete the fish stock size in the long run.

6.3 Future Work

This thesis is a comprehensive study of analysing the dynamical behaviour of fish populations, which could be extended for future research; below are some specific guidelines for further research.

6.3.1 Modelling a new Stock-recruitment relationship

One can probably look to overcome the limitation of the Deriso-Schnute model and find ways to interpolate between the Beverton-Holt, Ricker and Schaefer models without falling to the singularity at $\gamma = 0$.

6.3.2 Enhancing REMO

The main weakness of REMO is in predicting negative values for variables in response to ecological perturbations; one may enhance REMO by means of non-linear models so as to overcome the limitation caused by the linear regression models.

6.3.3 Dynamical Stability

An important ecological aspect that one may need to understand is whether there are ways to organise interaction of species so as to lead to more persistent communities. Scientists addressed this question by studying for example the relationship between diversity and stability (Ives and Carpenter, 2007) and the structural stability across ecological systems (Rudolf et al., 2014) so as to find ways for food web configurations that promote stable equilibrium population dynamics of species. Wilkinson (2011) defined an equilibrium solution as a set of concentrations that will not change over time; one can employ a stochastic process to approximate the evolution of the ecological system as a continuous time Markov process with a discrete state space. For instance, a stochastic kinetic formulation of the Lotka-Volterra model can be used to draw discrete event simulations using the Gillespie algorithm (Gillespie, 1977). A competitive Lotka-Volterra model can be described as follows:

$$\frac{dX_i}{dt} = r_i X_i \left(1 - \frac{\sum_{j=1}^{N} \alpha_{ij} X_j}{K_i}\right), \tag{6.1}$$

where N is the number of species, X is the size of the population at time t, r is the growth rate, and α_{ij} is the effect of species j on species i, which is also known as the trophic link strength; whereas K is the carrying capacity. We can use this model to simulate the equilibrium condition for the ecological system and understand the factors that may affect its stability. Moreover, we can ask a few important questions, such as:

- can we evaluate how robust communities are to species loss?
- what would happen if we eliminate some of the species in that region?

6.3.4 Autoregressive Hidden Markov Model

I propose employing a Bayesian state space model, also known as Autoregressive Hidden Markov Model, to predict the population trajectory of related species, given current stock assessment value and prior knowledge. This research aims to help us understand the evolution of marine species so as to evaluate the percentage of growth or decline of each particular species per year. Some ideas are already proposed for this task, which are described in Appendix G.

Appendix A

Populations Classification

Table A.1: Reliable fit of η_1 applied to the 90's S-R populations where we restrict the confidence level to 95%. The column 'Label' indicates whether there is a strong evidence for reliably identifying η_1, columns 2-to-6 report information about the populations, and γ indicates the best-fit model.

Label	Assessment Id	Method Short	Species	Common Name	Area Name	γ
-1	DFO-QUE-COD3Pn4RS-1964-2007	ADAPT	morhua	Atlantic cod	Northern Gulf of St. Lawrence	1
-1	IMARPE-PANCHPERUNC-1963-2004	VPA	ringens	Peruvian anchoveta	North-Central Peruvian coast	1
-1	INIDEP-SBWHITARGS-1985-2007	VPA	australis	Southern blue whiting	Southern Argentina	1
-1	NRIFS-OFLOUNECS-1986-2010	VPA	olivaceus	Olive flounder	East China Sea	1
-1	NWWG-HERRIsum-1984-2011	NFT-ADAPT	harengus	Herring	Iceland Grounds	0
-1	WGBFAS-HERR30-1972-2011	XSA	harengus	Herring	Bothnian Sea	1
-1	WGNSSK-WHITNS-VIId-IIIa-1989-2010	XSA	merlangus	Whiting	IIIa, VIId and North Sea	-2
0	AFWG-GHALNEAR-1960-2010	XSA	hippoglossoides	Greenland halibut	North-East Arctic	-1
0	AFWG-HADNEAR-1947-2010	XSA	aeglefinus	Haddock	North-East Arctic	1
0	AFWG-HADNS-IIIa-1963-2011	FLXSA	aeglefinus	Haddock	IIIa and North Sea	-1
0	AFWG-POLLNEAR-1957-2011	XSA	virens	Pollock	North-East Arctic	-1
0	DFO-COD5Zjm-1978-2003	ADAPT	morhua	Atlantic cod	Georges Bank	1
0	DFO-HAD5Zejm-1968-2003	ADAPT	aeglefinus	Haddock	Georges Bank	-1
0	DFO-HERR4VWX-1964-2006	ADAPT	harengus	Herring	Scotian Shelf and Bay of Fundy	0
0	DFO-MAR-HAD4X5Y-1960-2003	SPA-ADAPT	aeglefinus	Haddock	Western Scotian Shelf, Bay of Fundy and Gulf of Maine	0
0	DFO-NFLD-COD2J3KLIS-1959-2006	ADAPT	morhua	Atlantic cod	Southern Labrador-Eastern Newfoundland	0
0	DFO-NFLD-COD3Ps-1959-2004	B-ADAPT	morhua	Atlantic cod	St. Pierre Bank	1
0	DFO-POLL4X5YZ-1980-2006	ADAPT	virens	Pollock	Western Scotian Shelf, Bay of Fundy, Gulf of Maine and Georges Bank	-1
0	DFO-QUE-HERR4RFA-1971-2003	SPA-ADAPT	harengus	Herring	NAFO division 4R	0

0	DFO-QUE-HERR4RSP-1963-2004	SPA-ADAPT	harengus	Herring	NAFO division 4R	0
0	DFO-SG-COD4TVn-1965-2009	ADAPT	morhua	Atlantic cod	Southern Gulf of St. Lawrence	1
0	DFO-SG-HERR4TFA-1974-2007	SPA-ADAPT	harengus	Herring	Southern Gulf of St. Lawrence	0
0	DFO-SG-HERR4TSP-1974-2007	SPA-ADAPT	harengus	Herring	Southern Gulf of St. Lawrence	-1
0	HAWG-HERRVIaVIIbc-1956-2010	VPA	harengus	Herring	VIa, VIIb and VIIc	-1
0	ICCAT-ATBTUNAEATL-1950-2010	ADAPT	thynnus	Atlantic bluefin tuna	Eastern Atlantic	0
0	ICCAT-ATBTUNAWATL-1950-2010	ADAPT	thynnus	Atlantic bluefin tuna	Western Atlantic	-2
0	INIDEP-ARGANCHONARG-1989-2007	ADAPT	anchoita	Argentine anchoita	Northern Argentina	1
0	INIDEP-ARGHAKENARG-1985-2007	VPA	hubbsi	Argentine hake	Northern Argentina	0
0	INIDEP-ARGHAKESARG-1985-2008	VPA	hubbsi	Argentine hake	Southern Argentina	0
0	INIDEP-PATGRENADIERSARG-1983-2006	VPA	magellanicus	Patagonian grenadier	Southern Argentina	-1
0	NAFO-SC-AMPL3LNO-1955-2007	VPA	platessoides	American Plaice	Grand Banks	-2
0	NAFO-SC-AMPL3M-1960-2007	XSA	platessoides	American Plaice	Flemish Cap	1
0	NAFO-SC-COD3M-1959-2008	hybrid	morhua	Atlantic cod	Flemish Cap	1
0	NAFO-SC-COD3NO-1953-2007	SPA	morhua	Atlantic cod	Southern Grand Banks	1
0	NAFO-SC-GHAL23KLMNO-1960-2006	XSA	hippoglossoides	Greenland halibut	Labrador Shelf - Grand Banks	0
0	NAFO-SC-REDFISHSPP3M-1985-2006	XSA	spp	Redfish species	Flemish Cap	-1
0	NEFSC-AMPL5YZ-1960-2008	ADAPT	platessoides	American Plaice	Gulf of Maine / Georges Bank	-1
0	NEFSC-CODGB-1960-2008	ADAPT	morhua	Atlantic cod	Georges Bank	-2
0	NEFSC-CODGOM-1893-2008	ADAPT	morhua	Atlantic cod	Gulf of Maine	-1
0	NEFSC-HAD5Y-1956-2008	NFT-ADAPT	aeglefinus	Haddock	Gulf of Maine	-1
0	NEFSC-HADGB-1930-2008	NFT-ADAPT	aeglefinus	Haddock	Georges Bank	1
0	NEFSC-MACKGOMCHATT-1960-2005	VPA	scombrus	Mackerel	Gulf of Maine / Cape Hatteras	1
0	NEFSC-WINFLOUN5Z-1982-2007	ADAPT	americanus	Winter Flounder	Georges Bank	-1
0	NEFSC-WINFLOUNSNEMATL-1940-2007	NFT-ADAPT	americanus	Winter Flounder	Southern New England /Mid Atlantic	1
0	NEFSC-WITFLOUN5Y-1982-2008	VPA	cynoglossus	Witch Flounder	Gulf of Maine	-1

0	NEFSC-YELLCCODGOM-1935-2008	VPA	ferruginea	Yellowtail flounder	Cape Cod / Gulf of Maine	-1
0	NEFSC-YELLGB-1935-2008	VPA	ferruginea	Yellowtail flounder	Georges Bank	0
0	NEFSC-YELLSNEMATL-1935-2008	VPA	ferruginea	Yellowtail Flounder	Southern New England /Mid Atlantic	0
0	NRIFS-BMACKECS-1992-2010	VPA	australasicus	Blue mackerel	East China Sea	-2
0	NRIFS-CMACKTSST-1973-2010	VPA	japonicus	Chub mackerel	Tsushima Strait	1
0	NRIFS-JANCHOPJPN-1978-2009	VPA	japonicus	Japanese anchovy	Pacific Coast of Japan	-2
0	NRIFS-JMACKTSST-1973-2010	VPA	japonicus	Japanese jack mackerel	Tsushima Strait	1
0	NRIFS-OFLOUNNSJ-1999-2010	VPA	olivaceus	Olive flounder	Sea of Japan North	-2
0	NRIFS-OFLOUNSETO-1987-2010	VPA	olivaceus	Olive flounder	Inland Sea of Japan	-1
0	NRIFS-PILCHTSST-1960-2010	VPA	melanostictus	Japanese pilchard	Tsushima Strait	1
0	NRIFS-RBRMPAC-1977-2010	VPA	major	Red seabream	Pacific Ocean	-1
0	NRIFS-SAURNWPAC-1980-2010	VPA	saira	Pacific saury	Northwest Pacific	-2
0	NRIFS-SPANMACKSETO-1987-2010	VPA	niphonius	Japanese Spanish mackerel	Inland Sea of Japan	1
0	NWWG-CODFAPL-1959-2011	XSA	morhua	Atlantic cod	Faroe Plateau	-1
0	NWWG-HADFAPL-1955-2011	XSA	aeglefinus	Haddock	Faroe Plateau	1
0	NWWG-HADICE-1977-2011	XSA	aeglefinus	Haddock	Iceland Grounds	0
0	NWWG-POLLFAPL-1958-2011	XSA	virens	Pollock	Faroe Plateau	1
0	SEFSC-KMACKGM-1992-2001	VPA	cavalla	King Mackerel	Gulf of Mexico	0
0	SEFSC-KMACKSATLC-1981-2001	VPA	cavalla	King Mackerel	Southern Atlantic coast	1
0	WGBFAS-CODIS-1968-2010	B-ADAPT	morhua	Atlantic cod	Irish Sea	1
0	WGBFAS-CODVIIek-1970-2011	XSA	morhua	Atlantic Cod	Celtic Sea	1
0	WGBFAS-HERR2532-1973-2011	XSA	harengus	Herring	Eastern Baltic	-2
0	WGBFAS-HERR31-1979-2010	XSA	harengus	Herring	Bothnian Bay	1
0	WGBFAS-HERRRIGA-1976-2011	XSA	harengus	Herring	Gulf of Riga East of Gotland	1
0	WGBFAS-SPRAT22-32-1973-2011	XSA	sprattus	Sprat	Baltic Areas 22-32	-2
0	WGHMM-FMEG8c9a-1986-2010	XSA	boscii	Fourspotted megrim	VIIIc-IXa	-2
0	WGHMM-MEG8c9a-1985-2010	XSA	whiffiagonis	Megrim	VIIIc-IXa	1
0	WGHMM-SOLEVIII-1982-2011	XSA	vulgaris	common European sole	Bay of Biscay	1
0	WGNSDS-SOLEIS-1968-2011	XSA	vulgaris	common European sole	Irish Sea	1
0	WGNSSK-CODCOASTNOR-1982-2010	XSA	morhua	Atlantic cod	North-East Arctic	1
0	WGNSSK-CODNEAR-1943-2010	XSA	morhua	Atlantic cod	North-East Arctic	0

0	WGNSSK-HADROCK-1990-2011	XSA	aeglefinus	Haddock	Rockall Bank	-1
0	WGNSSK-NPOUTNS-1983-2011	SXSA	esmarkii	Norway pout	North Sea	-1
0	WGNSSK-PLAIC7d-1979-2010	XSA	platessa	European Plaice	Eastern English Channel	1
0	WGNSSK-PLAICIIIa-1976-2006	XSA	platessa	European Plaice	Kattegat and Skagerrak	0
0	WGNSSK-PLAICNS-1956-2010	XSA	platessa	European Plaice	North Sea	-1
0	WGNSSK-POLLNS-VI-IIIa-1964-2010	XSA	virens	Pollock	IIIa, VI and North Sea	-1
0	WGNSSK-SOLENS-1956-2010	XSA	vulgaris	common European sole	North Sea	0
0	WGNSSK-SOLEVIId-1981-2011	XSA	vulgaris	common European sole	Eastern English Channel	-1
0	WGSSDS-HADVIIb-k-1993-2006	XSA	aeglefinus	Haddock	ICES VIIb-k	-1
0	WGSSDS-PLAICECHW-1979-2010	XSA	platessa	European Plaice	Western English Channel	1
0	WGSSDS-SOLECS-1970-2011	XSA	vulgaris	common European sole	Celtic Sea	-1
0	WGSSDS-SOLEVIIe-1968-2010	FLXSA	vulgaris	common European sole	Western English Channel	-2
0	WGSSDS-WHITVIIek-1982-2010	FLXSA	merlangus	Whiting	Celtic Sea	-1
1	NRIFS-RBRMSETO-1977-2010	VPA	major	Red seabream	Inland Sea of Japan	0

Appendix B

Derivative of Expected Fish Recruitment

The derivative of expected fish recruitment starts by considering Equation (3.3), such that

$$\ln\left(\frac{R_i}{S_i}\right) \sim_{\text{i.i.d.}} \mathcal{N}(\mu_i, \sigma_i^2) \quad \text{where} \quad \mu_i = \ln(\alpha) + \frac{1}{\gamma}\ln(1-\gamma\beta S_i) \quad \text{and}$$
$$\sigma_i^2 = \exp(\eta_0 + \eta_1 S_i),$$

By the properties of the lognormal distribution, the expectation of $\ln(R/S)$ can be described as

$$\mathbb{E}(R/S) = \exp(\mu + \sigma^2/2).$$

For a fixed S, the expected recruits becomes

$$\mathbb{E}(R) = S\exp\left(\ln(\alpha) + \frac{1}{\gamma}\ln(1-\gamma\beta S) + \exp(\eta_0 + \eta_1 S)/2\right). \tag{B.1}$$

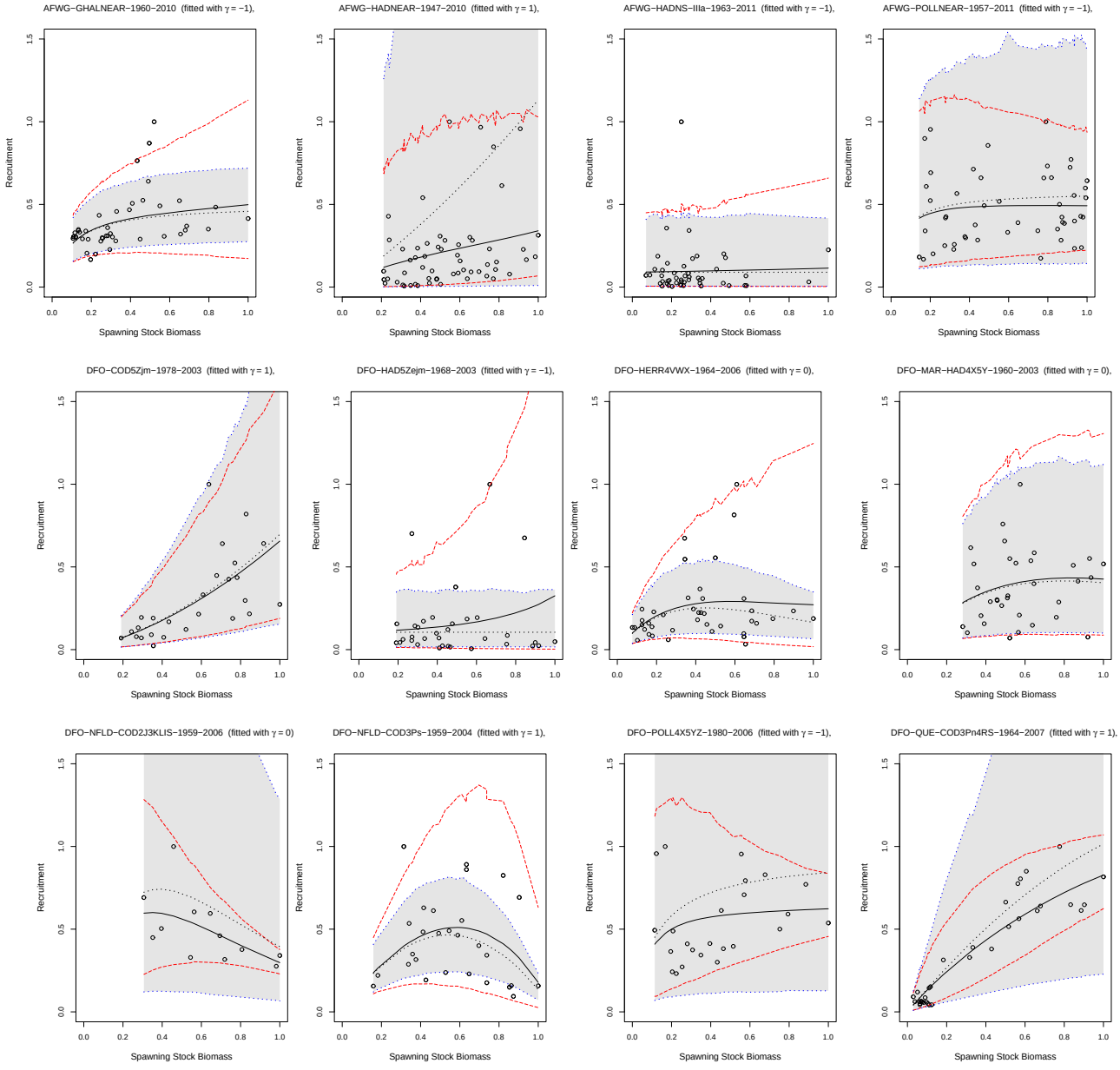
For $\gamma = -2$, then

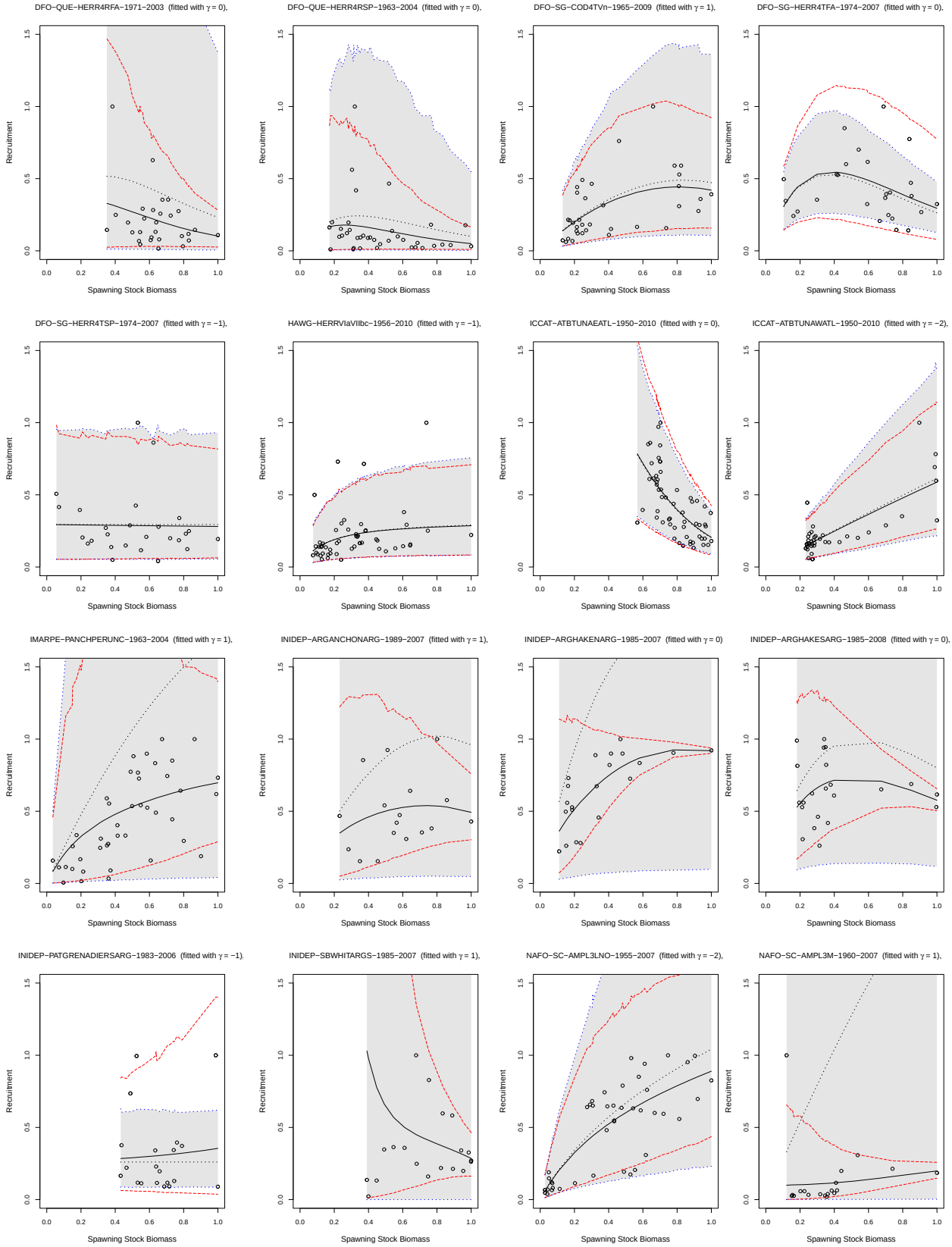
$$\begin{aligned}
\mathbb{E}(R) &= S\exp\left(\ln(\alpha) - \frac{1}{2}\ln(1+2\beta S) + \exp(\eta_0 + \eta_1 S)/2\right) \\
&= S\exp\left(\ln(\alpha)\right)\exp\left(-\frac{1}{2}\ln(1+2\beta S)\right)\exp\left(\exp(\eta_0 + \eta_1 S)/2\right) \\
&= \frac{\alpha S}{\sqrt{1+2\beta S}}\exp\left(\frac{\exp(\eta_0 + \eta_1 S)}{2}\right).
\end{aligned}$$

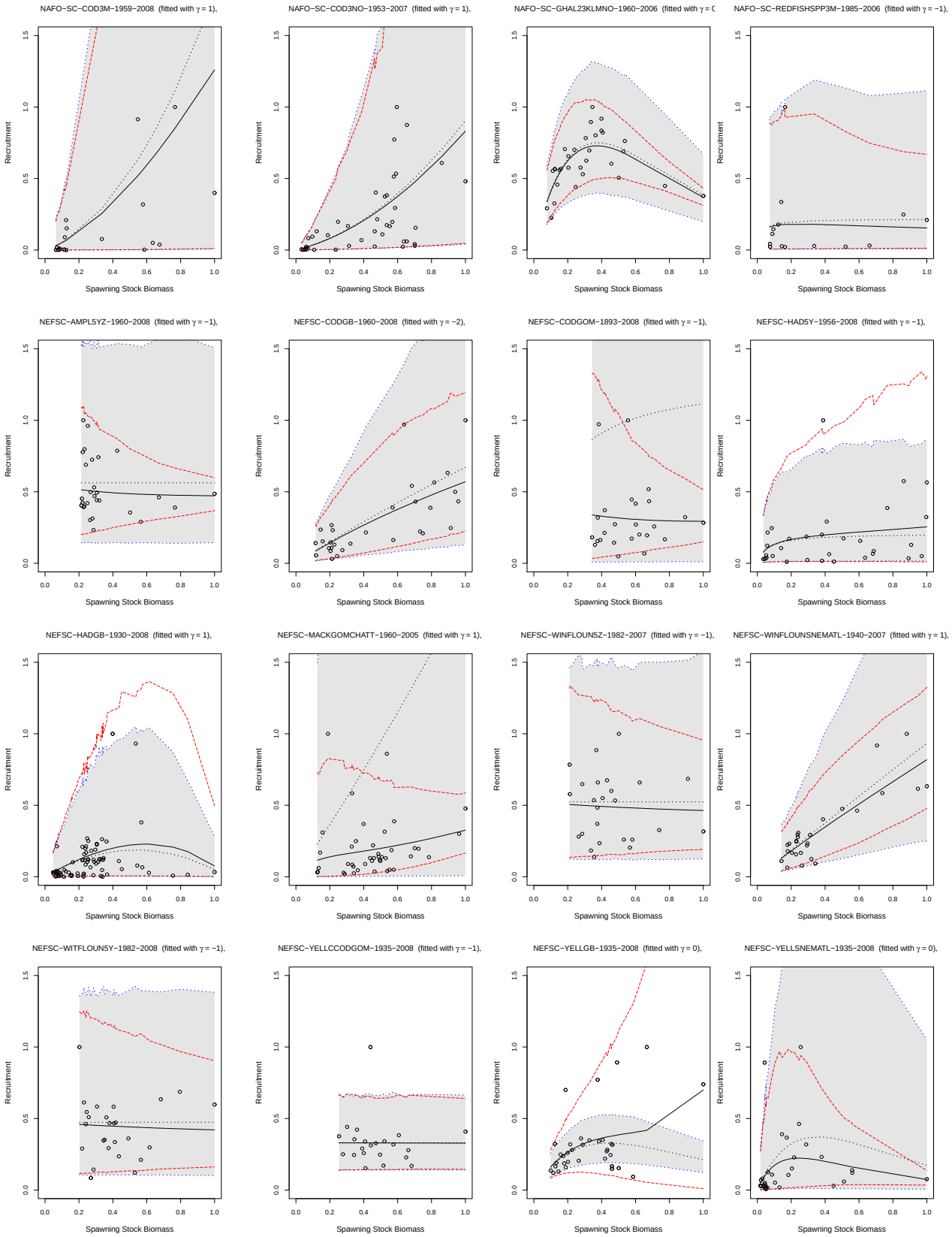
Appendix C

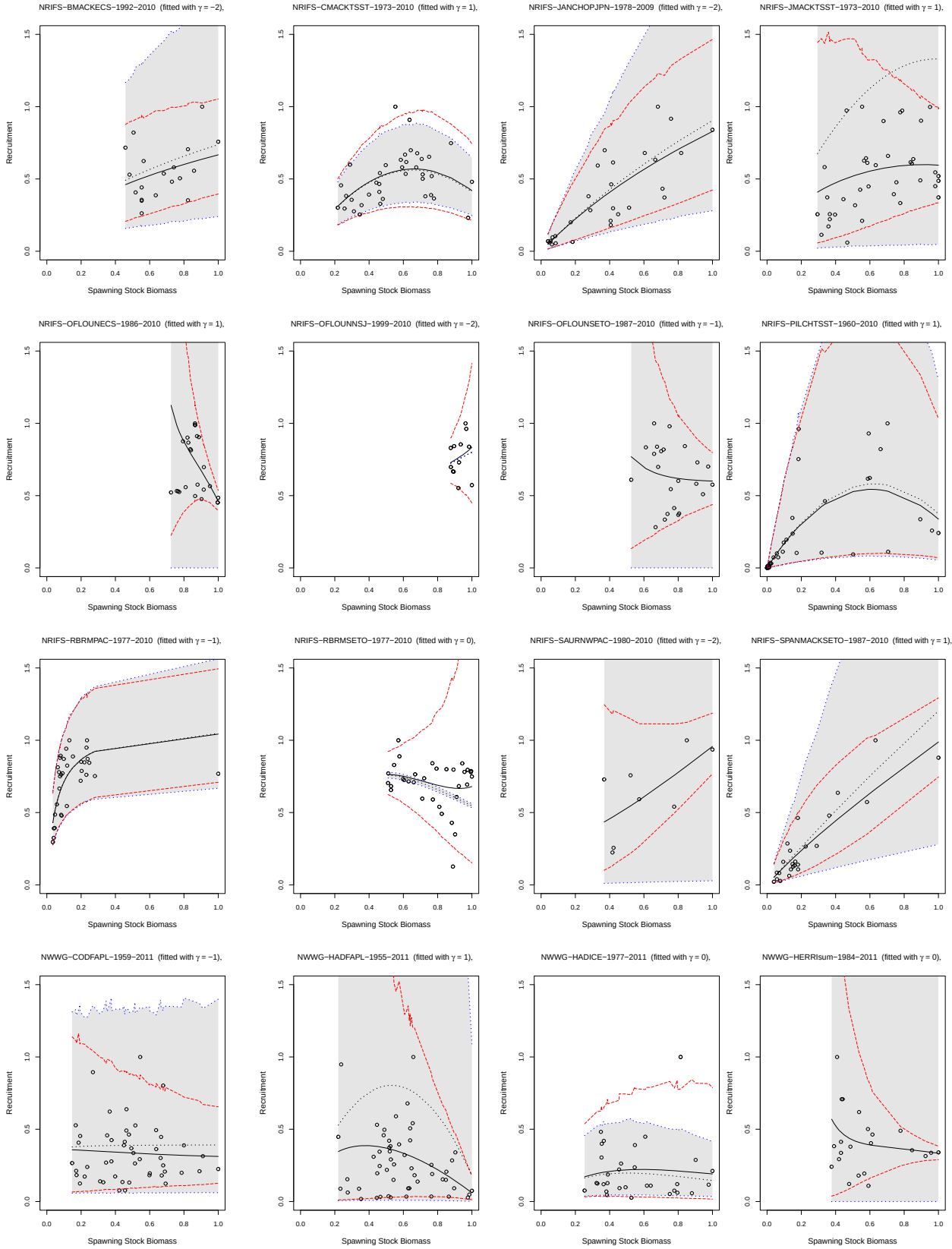
Expected stock and recruitment curves

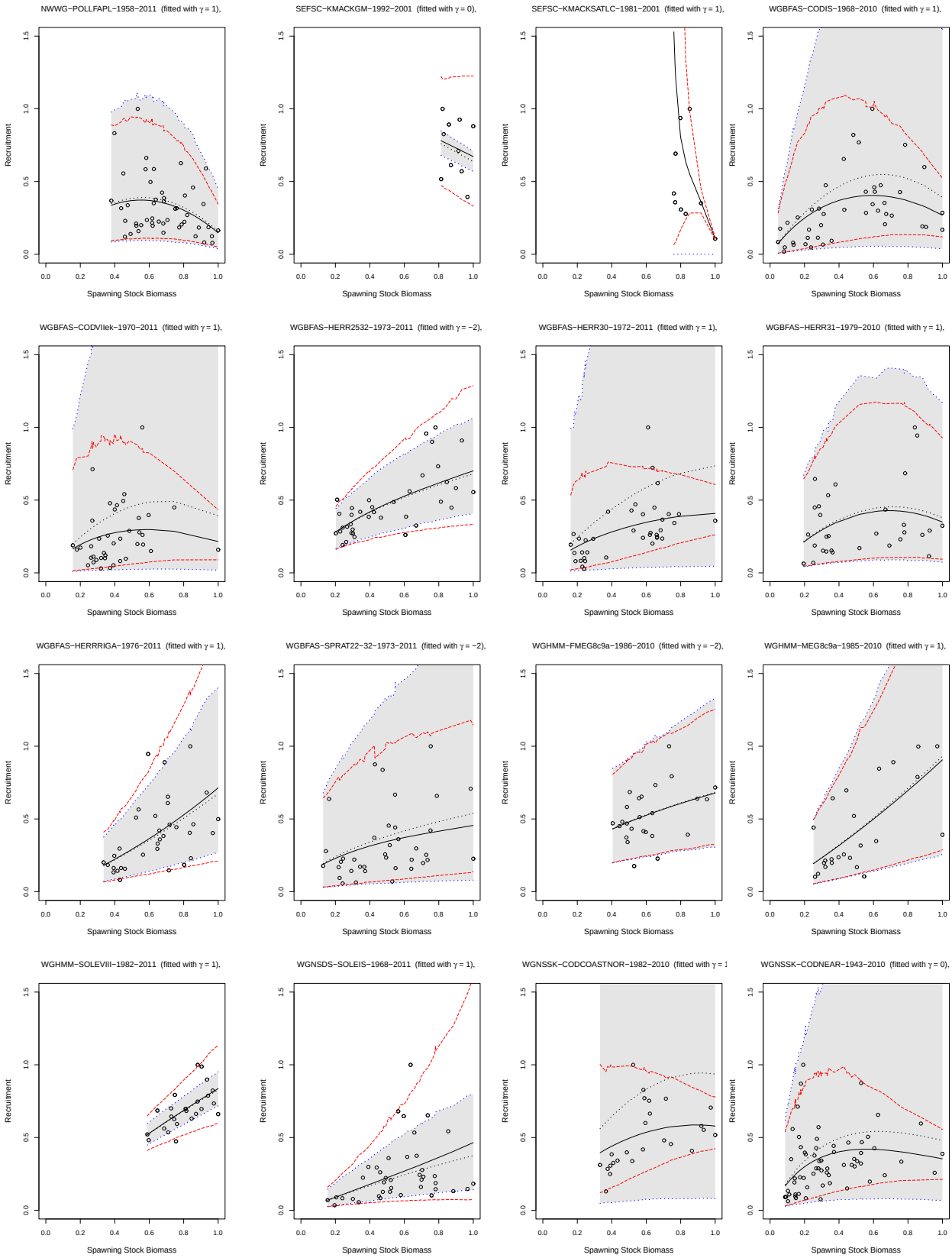
Figure C.1: Expected stock-recruitment curves with approximate 95% confidence intervals fitted with different values of γ. Examples of the 90 S-R datasets that illustrate the difference in fit between the heteroscedastic and nonheteroscedastic models. The expected recruit for the nonheteroscedastic model (dotted black plot) and its approximate 95% confidence interval (grey area) are compared against the expected recruit for the heteroscedastic model (solid black plot) and its approximated 95% confidence interval (dashed plot).

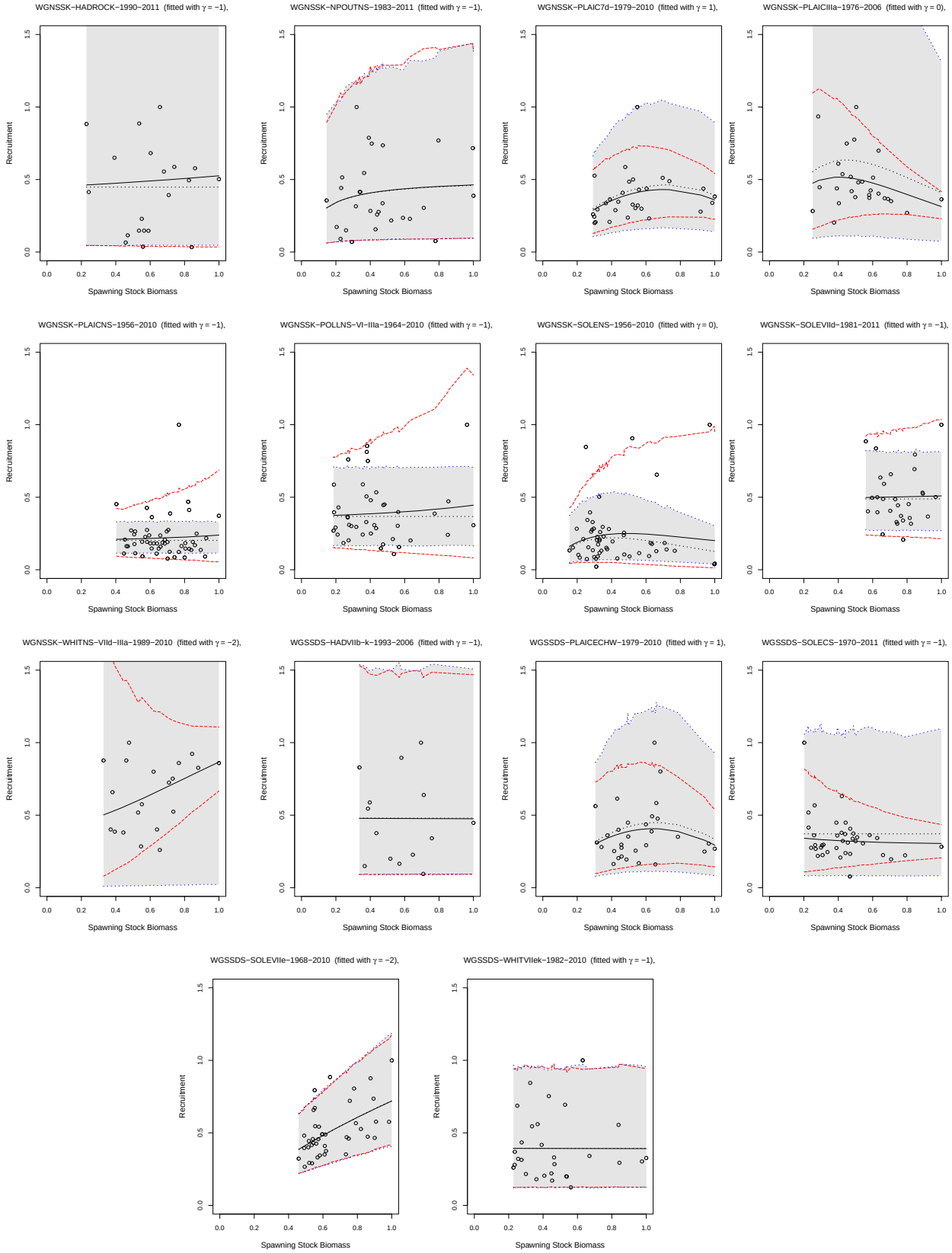



DFO−QUE−HERR4RFA−1971−2003 (fitted with γ = 0),
DFO−QUE−HERR4RSP−1963−2004 (fitted with γ = 0),
DFO−SG−COD4TVn−1965−2009 (fitted with γ = 1),
DFO−SG−HERR4TFA−1974−2007 (fitted with γ = 0),
DFO−SG−HERR4TSP−1974−2007 (fitted with γ = −1),
HAWG−HERRVIaVIIbc−1956−2010 (fitted with γ = −1),
ICCAT−ATBTUNAEATL−1950−2010 (fitted with γ = 0),
ICCAT−ATBTUNAWATL−1950−2010 (fitted with γ = −2),
IMARPE−PANCHPERUNC−1963−2004 (fitted with γ = 1),
INIDEP−ARGANCHONARG−1989−2007 (fitted with γ = 1),
INIDEP−ARGHAKENARG−1985−2007 (fitted with γ = 0)
INIDEP−ARGHAKESARG−1985−2008 (fitted with γ = 0),
INIDEP−PATGRENADIERSARG−1983−2006 (fitted with γ = −1),
INIDEP−SBWHITARGS−1985−2007 (fitted with γ = 1),
NAFO−SC−AMPL3LNO−1955−2007 (fitted with γ = −2),
NAFO−SC−AMPL3M−1960−2007 (fitted with γ = 1),
Recruitment
Spawning Stock Biomass


NAFO−SC−COD3M−1959−2008 (fitted with γ = 1),
NAFO−SC−COD3NO−1953−2007 (fitted with γ = 1),
NAFO−SC−GHAL23KLMNO−1960−2006 (fitted with γ = 0
NAFO−SC−REDFISHSPP3M−1985−2006 (fitted with γ = −1),
NEFSC−AMPL5YZ−1960−2008 (fitted with γ = −1),
NEFSC−CODGB−1960−2008 (fitted with γ = −2),
NEFSC−CODGOM−1893−2008 (fitted with γ = −1),
NEFSC−HAD5Y−1956−2008 (fitted with γ = −1),
NEFSC−HADGB−1930−2008 (fitted with γ = 1),
NEFSC−MACKGOMCHATT−1960−2005 (fitted with γ = 1),
NEFSC−WINFLOUN5Z−1982−2007 (fitted with γ = −1),
NEFSC−WINFLOUNSNEMATL−1940−2007 (fitted with γ = 1),
NEFSC−WITFLOUN5Y−1982−2008 (fitted with γ = −1),
NEFSC−YELLCCODGOM−1935−2008 (fitted with γ = −1),
NEFSC−YELLGB−1935−2008 (fitted with γ = 0),
NEFSC−YELLSNEMATL−1935−2008 (fitted with γ = 0),
Recruitment
Spawning Stock Biomass


NRIFS−BMACKECS−1992−2010 (fitted with γ = −2),
NRIFS−CMACKTSST−1973−2010 (fitted with γ = 1),
NRIFS−JANCHOPJPN−1978−2009 (fitted with γ = −2),
NRIFS−JMACKTSST−1973−2010 (fitted with γ = 1),
NRIFS−OFLOUNECS−1986−2010 (fitted with γ = 1),
NRIFS−OFLOUNNSJ−1999−2010 (fitted with γ = −2),
NRIFS−OFLOUNSETO−1987−2010 (fitted with γ = −1),
NRIFS−PILCHTSST−1960−2010 (fitted with γ = 1),
NRIFS−RBRMPAC−1977−2010 (fitted with γ = −1),
NRIFS−RBRMSETO−1977−2010 (fitted with γ = 0),
NRIFS−SAURNWPAC−1980−2010 (fitted with γ = −2),
NRIFS−SPANMACKSETO−1987−2010 (fitted with γ = 1),
NWWG−CODFAPL−1959−2011 (fitted with γ = −1),
NWWG−HADFAPL−1955−2011 (fitted with γ = 1),
NWWG−HADICE−1977−2011 (fitted with γ = 0),
NWWG−HERRIsum−1984−2011 (fitted with γ = 0),
Recruitment
Spawning Stock Biomass


NWWG−POLLFAPL−1958−2011 (fitted with γ = 1),
SEFSC−KMACKGM−1992−2001 (fitted with γ = 0),
SEFSC−KMACKSATLC−1981−2001 (fitted with γ = 1),
WGBFAS−CODIS−1968−2010 (fitted with γ = 1),
WGBFAS−CODVIIek−1970−2011 (fitted with γ = 1),
WGBFAS−HERR2532−1973−2011 (fitted with γ = −2),
WGBFAS−HERR30−1972−2011 (fitted with γ = 1),
WGBFAS−HERR31−1979−2010 (fitted with γ = 1),
WGBFAS−HERRRIGA−1976−2011 (fitted with γ = 1),
WGBFAS−SPRAT22−32−1973−2011 (fitted with γ = −2),
WGHMM−FMEG8c9a−1986−2010 (fitted with γ = −2),
WGHMM−MEG8c9a−1985−2010 (fitted with γ = 1),
WGHMM−SOLEVIII−1982−2011 (fitted with γ = 1),
WGNSDS−SOLEIS−1968−2011 (fitted with γ = 1),
WGNSSK−CODCOASTNOR−1982−2010 (fitted with γ = 1
WGNSSK−CODNEAR−1943−2010 (fitted with γ = 0),
Recruitment
Spawning Stock Biomass


WGNSSK−HADROCK−1990−2011 (fitted with γ = −1),
WGNSSK−NPOUTNS−1983−2011 (fitted with γ = −1),
WGNSSK−PLAIC7d−1979−2010 (fitted with γ = 1),
WGNSSK−PLAICIIIa−1976−2006 (fitted with γ = 0),
WGNSSK−PLAICNS−1956−2010 (fitted with γ = −1),
WGNSSK−POLLNS−VI−IIIa−1964−2010 (fitted with γ = −1),
WGNSSK−SOLENS−1956−2010 (fitted with γ = 0),
WGNSSK−SOLEVIId−1981−2011 (fitted with γ = −1),
WGNSSK−WHITNS−VIId−IIIa−1989−2010 (fitted with γ = −2),
WGSSDS−HADVIIb−k−1993−2006 (fitted with γ = −1),
WGSSDS−PLAICECHW−1979−2010 (fitted with γ = 1),
WGSSDS−SOLECS−1970−2011 (fitted with γ = −1),
WGSSDS−SOLEVIIe−1968−2010 (fitted with γ = −2),
WGSSDS−WHITVIIek−1982−2010 (fitted with γ = −1),
Recruitment
Spawning Stock Biomass

Appendix D

Marginal Likelihood Estimation

Here I include the implemented JAGS model used to estimate the marginal likelihood for the model $\mathcal{M}_1$ (with $\gamma < 0$) via the power posterior method, as described in Chapter 4.

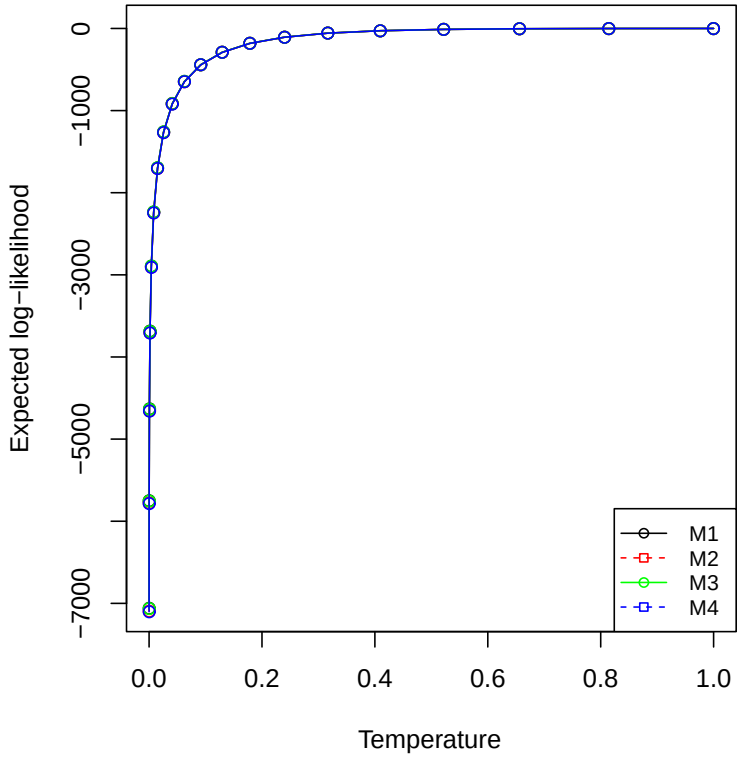
```
data {
  for( j in 1 : L ) { # j denotes different populations, L is a parameter.
     for(i in StartFrom.st[1,j] : StopAt.st[1,j]) { # i denotes datapoints, 'StartFrom
         .st' and 'StopAt.st' are vectors indicating the size of each population (i.e. a
         ll fish populations were put in a matrix in R).
       for (s in 1:T){ # s denotes the cutpoints.
         log.RdivS[i,j,s] <- (log(R[i,j]/S[i,j]))
  }} }
 }

 model {
 for(j in 1 : L ) {
   for(s in 1 : T){
     for( i in StartFrom.st[1,j] : StopAt.st[1,j]) {
       # model's likelihood.
       log.RdivS[i,j,s] ~ dnorm(mu[i,j,s],tau[i,j,s])
       mu[i,j,s]        <- (log.alpha[j,s] + (1/gamma[j,s])*log(1 - gamma[j,s]*beta[j,s]
           *S[i,j]))
       tau[i,j,s]       <-  exp(-eta0[j,s] - eta1[j,s]*S[i,j])

       # I divide the likelihood in four terms, for simplification.
       termA[i,j,s] <- 0.5*N.data[j]*log(2*PI)
       termB[i,j,s] <- 0.5*sum(eta0[j,s] + eta1[j,s]*S[i,j])
       termC[i,j,s] <- (log.RdivS[i,j,s] - log.alpha[j,s] - (1/gamma[j,s])*log(1-gamma[j
           ,s]*beta[j,s]*S[i,j]))^2
       termD[i,j,s] <- exp(eta0[j,s] + eta1[j,s]*S[i,j])
       # by tempering the log-likelihood expression, we write
       logLikelihood[i,j,s]  <- (-termA[i,j,s] - termB[i,j,s] -0.5*sum(termC[i,j,s]/term
           D[i,j,s]))*q[T+1-s]
      }# end i.
```

```
    # sum up the Log-Likelihood of all datapoints belonging to each population respect
        ively.
    LLik[j,s] <-  sum(logLikelihood[StartFrom.st[1,j]:StopAt.st[1,j],j,s])
   }#end s.
  }#end j.

  for(s in 1 : T){
   for(j in 1 : L ) {
    # define the random effects.
    log.alpha[j,s]  ~  dnorm(phi.a[s],  1/pow(0.1,2))
    beta[j,s]       ~  dgamma(phi.k.b[s], 0.01)
    eta0[j,s]       ~  dnorm(phi.e0[s], 1/pow(0.4,2))
    eta1[j,s]       ~  dnorm(phi.e1[s], 1/pow(1.5,2))
    gamma[j,s]      ~  dnorm(phi.g[s],  1/pow(0.2,2))
   }#end j.

   # define the hyperprior distributions.
   phi.a[s]   ~ dunif(-5, 20)       #-->alpha.
   phi.k.b[s] ~ dunif(1,300)        #-->beta.
   phi.e0[s]  ~ dunif(-20,20)       #-->eta0.
   phi.e1[s]  ~ dunif(-20,20)       #-->eta1.
   phi.g[s]   ~ dunif(-10, -1)      #-->gamma.
   #..
   q[s] <-pow(a[s],e)     # 'e': is the exponent parameter set from the main script (e.g
       . varies from 2 -to- 5).
   # estimate the log-likelihood at different cut-points.
   tLL[s] <- sum(LLik[1:L, s])
  }#end s

  PI <- 3.1415926
  a[1] <- 0.01
  for(s in 2:T)  {a[s] <- s/T}
  for(s in 1:(T-1)) {
   # compute the log of the evidence using the trapezium rule.
   mc.LL[s] <-  (q[s+1]-q[s])*(tLL[s+1]+tLL[s])*0.5
  }
  # get the sum of all Monte-Carlo (mc) cut-point samples.
  logML <- sum(mc.LL[])
}
```

Figure D.1 plots the expected deviances $\mathbf{E}_{\boldsymbol{\theta}|\ln(\boldsymbol{R}/\boldsymbol{S}),\mathbf{t}}[\log\{p(\ln(\boldsymbol{R}/\boldsymbol{S})|\boldsymbol{\theta})\}]$ for the power posterior for the different models $\mathcal{M}_1$, $\mathcal{M}_2$, $\mathcal{M}_3$ and $\mathcal{M}_4$ with $\gamma < 0$ against t. The analysis is based on the North-East Arctic containing five populations and where the model is examined over 20 discretized step $(0 = t_0 < t_1 < \ldots t_{19} < t_{20} = 1)$ using a temperature schedule of the type $t_i = (i/20)^4$. To improve the convergence rate, I started at t_{20} and used the last posterior mean parameter values from the MCMC chain to initiate the chain at the previous temperature step t_{19} and so forth. The analysis is executed over four parallel MCMC chains

where after burn-in and thinning I collect 4000 samples from the stationary distribution $p_{t_s}\{\boldsymbol{\theta}|\ln(\boldsymbol{R}/\boldsymbol{S})\}$. This plot shows that there is a little differences in $\mathbf{E}_{\boldsymbol{\theta}|\ln(\boldsymbol{R}/\boldsymbol{S}),\mathbf{t}}[\log\{p(\ln(\boldsymbol{R}/\boldsymbol{S})|\boldsymbol{\theta})\}]$ for t away from 0.

Figure D.1: Expected (half) deviance under the distribution $p_t\{\boldsymbol{\theta}|\ln(\boldsymbol{R}/\boldsymbol{S})\}$, plotted against temperature for the models: $\mathcal{M}_1$, $\mathcal{M}_2$, $\mathcal{M}_3$ and $\mathcal{M}_4$ (with $\gamma < 0$) applied to the North-East Arctic region.



Appendix E

Recruitment Prediction in JAGS

Here I include the source code of the JAGS model used to predict the marginal posterior distribution of fish recruitment for each test data point using the model $\mathcal{M}_1$ with $\gamma < 0$.

```
data {
  for( j in 1 : L ) {
    for(i in StartFrom.st[1,j] : StopAt.Train.st[1,j]) {
      log.RdivS[i,j] <- log(R[i,j]/S[i,j])
    } }
}

model {
  for(j in 1 : L ) {
    for( i in StartFrom.st[1,j] : StopAt.Train.st [1,j]) {
      # model's likelihood
      log.RdivS[i,j]  ~ dnorm(mu[i,j],tau[i,j])
      mu[i,j]        <-  log.alpha[j] + (1/gamma[j])*log(1 - gamma[j]*beta[j]*S[i,j])
      tau[i,j]       <-  exp(-eta0[j] - eta1[j]*S[i,j])
    }
    # random effects
    log.alpha[j]  ~  dnorm(phi.a,  1/pow(0.2,2))
    beta[j]       ~  dgamma(phi.k.b, 0.01)
    eta0[j]       ~  dnorm(phi.e0, 1/pow(0.4,2))
    eta1[j]       ~  dnorm(phi.e1, 1/pow(1.5,2))
    gamma[j]      ~  dnorm(phi.g,  1/pow(0.2,2))
  }
  # Hyperprior distributions
  phi.a   ~ dunif(-5, 20)          #-->alpha
  phi.k.b ~ dunif(1,300)           #-->beta
  phi.e0  ~ dunif(-20,20)          #-->eta0
  phi.e1  ~ dunif(-20,20)          #-->eta1
  phi.g   ~ dunif(-10, -1)         #-->gamma

 #do recruitment prediction.
```

```
   for(j in 1 : L ) {
    for( i in StartFrom.Test.st[1,j] : StopAt.st[1,j]) {
        # model's likelihood
        log.RdivS.Test[i,j]  ~ dnorm(mu.test[i,j],tau.test[i,j])
        mu.test[i,j]         <-  log.alpha[j] + (1/gamma[j])*log(1 - gamma[j]*beta[j]*S[i,
            j])
        tau.test[i,j]        <-  exp(-eta0[j] - eta1[j]*S[i,j])

        R.pred[i,j] <- exp(log.RdivS.Test[i,j])*S[i,j]
    }
       }
}
```

Appendix F

Dynamic Bayesian Network Learning

F.1 Proposed model with listwise deletion for fish populations (REMO1)

I define explicitly the set of interactions of the Proposed model in term of multivariate first order autoregressive model as per below:

$$
\begin{aligned}
X_{\mathrm{AO}}(t+1) &= \alpha_1 X_{\mathrm{SOI}}(t). \\
X_{\mathrm{NAO}}(t+1) &= \alpha_2 X_{\mathrm{SOI}}(t). \\
X_{\mathrm{SST}}(t+1) &= \alpha_3 X_{\mathrm{SST}}(t) + \alpha_4 X_{\mathrm{SAL}}(t). \\
X_{\mathrm{Wind}}(t+1) &= \alpha_5 X_{\mathrm{Krill}}(t) + \alpha_6 X_{\mathrm{SAL}}(t). \\
X_{\mathrm{SAL}}(t+1) &= \alpha_7 X_{\mathrm{SAL}}(t) + \alpha_8 X_{\mathrm{Diatom}}(t) + \alpha_9 X_{\mathrm{SST}}(t). \\
X_{\mathrm{SOI}}(t+1) &= \alpha_{10} X_{\mathrm{SOI}}(t). \\
X_{\mathrm{FishLarvae}}(t+1) &= \alpha_{11} X_{\mathrm{SST}}(t) + \alpha_{12} X_{\mathrm{Cope}}(t). \\
X_{\mathrm{Krill}}(t+1) &= \alpha_{13} X_{\mathrm{SST}}(t) + \alpha_{14} X_{\mathrm{Diatom}}(t) + \alpha_{15} X_{\mathrm{FishLarvae}}(t) + \\
&\quad \alpha_{16} X_{\mathrm{SAL}}(t) + \alpha_{17} X_{\mathrm{SST}}(t) X_{\mathrm{FishLarvae}}(t). \\
X_{\mathrm{LargeCope}}(t+1) &= \alpha_{18} X_{\mathrm{Cope}}(t) + \alpha_{19} X_{\mathrm{SST}}(t).
\end{aligned}
$$

$$
\begin{aligned}
X_{\text{FishEgg}}(t+1) &= \alpha_{20}X_{\text{Cope}}(t) + \alpha_{21}X_{\text{Haddock}}(t) + \alpha_{22}X_{\text{AO}}(t) + \alpha_{23}X_{\text{Krill}}(t) + \\
&\quad \alpha_{24}X_{\text{NAO}}(t) + \alpha_{25}X_{\text{SST}}(t) + \alpha_{26}X_{\text{Haddock}}(t)X_{\text{AO}}(t). \\
X_{\text{Dinoflage}}(t+1) &= \alpha_{27}X_{\text{Dinoflage}}(t) + C. \\
X_{\text{Diatom}}(t+1) &= \alpha_{28}X_{\text{NAO}}(t) + \alpha_{29}X_{\text{Dinoflage}}(t) + C. \\
X_{\text{Cope}}(t+1) &= \alpha_{30}X_{\text{Megrim}}(t) + \alpha_{31}X_{\text{Diatom}}(t) + \alpha_{32}X_{\text{FishLarvae}}(t) + \alpha_{33}X_{\text{SST}}(t). \\
X_{\text{Cod}}(t+1) &= \alpha_{34}X_{\text{Cod}}(t) + \alpha_{35}X_{\text{fmortCod}}(t). \\
X_{\text{Haddock}}(t+1) &= \alpha_{36}X_{\text{Haddock}}(t) + \alpha_{37}X_{\text{fmortHad}}(t). \\
X_{\text{Herring}}(t+1) &= \alpha_{38}X_{\text{Herring}}(t) + \alpha_{39}X_{\text{fmortHer}}(t). \\
X_{\text{Megrim}}(t+1) &= \alpha_{40}X_{\text{Megrim}}(t) + \alpha_{41}X_{\text{fmortMeg}}(t). \\
X_{\text{Whiting}}(t+1) &= \alpha_{42}X_{\text{Whiting}}(t) + \alpha_{43}X_{\text{fmortWhiting}}(t).
\end{aligned}
$$

F.2 Proposed model with Biotic and Abiotic varilables (REMO2)

I define explicitly the set of interactions of the Proposed model (Biotic and Abiotic varilables only) in term of multivariate first order autoregressive model as per below:

$$
\begin{aligned}
X_{\text{AO}}(t+1) &= 0. \\
X_{\text{NAO}}(t+1) &= 0. \\
X_{\text{SST}}(t+1) &= \alpha_{1}X_{\text{SST}}(t). \\
X_{\text{Wind}}(t+1) &= \alpha_{2}X_{\text{SAL}}(t) + \alpha_{3}X_{\text{SST}}(t) + \alpha_{4}X_{\text{SOI}}(t) + \alpha_{5}X_{\text{Dinoflage}}(t) + \\
&\quad \alpha_{6}X_{\text{SAL}}(t)X_{\text{SST}}(t). \\
X_{\text{SAL}}(t+1) &= \alpha_{7}X_{\text{SAL}}(t). \\
X_{\text{SOI}}(t+1) &= 0. \\
X_{\text{FishLarvae}}(t+1) &= \alpha_{8}X_{\text{SST}}(t) + \alpha_{9}X_{\text{SAL}}(t) + \alpha_{10}X_{\text{Cope}}(t) + \alpha_{11}X_{\text{Dinoflage}}(t) + \\
&\quad \alpha_{12}X_{\text{LargeCope}}(t) + \alpha_{13}X_{\text{Diatom}}(t) + \alpha_{14}X_{\text{SST}}(t)X_{\text{Diatom}}(t).
\end{aligned}
$$

$$
\begin{aligned}
X_{\text{Krill}}(t+1) &= \alpha_{15} X_{\text{SST}}(t) + \alpha_{16} X_{\text{Cope}}(t) + \alpha_{17} X_{\text{FishLarvae}}(t). \\
X_{\text{LargeCope}}(t+1) &= \alpha_{18} X_{\text{Cope}}(t) + \alpha_{19} X_{\text{SST}}(t). \\
X_{\text{FishEgg}}(t+1) &= \alpha_{20} X_{\text{SAL}}(t) + \alpha_{21} X_{\text{SST}}(t) + \alpha_{22} X_{\text{SOI}}(t) + \alpha_{23} X_{\text{LargeCope}}(t) + \\
&\quad \alpha_{24} X_{\text{Diatom}}(t). \\
X_{\text{Dinoflage}}(t+1) &= \alpha_{25} X_{\text{SOI}}(t) + \alpha_{26} X_{\text{SST}}(t) + \alpha_{27} X_{\text{Wind}}(t) + \alpha_{28} X_{\text{NAO}}(t) + \\
&\quad \alpha_{29} X_{\text{Diatom}}(t) + \alpha_{30} X_{\text{SST}}(t) X_{\text{Diatom}}(t). \\
X_{\text{Diatom}}(t+1) &= \alpha_{31} X_{\text{Dinoflage}}(t) + \alpha_{32} X_{\text{SST}}(t) + \alpha_{33} X_{\text{FishLarvae}}(t) + \\
&\quad \alpha_{34} X_{\text{Wind}}(t) + \alpha_{35} X_{\text{NAO}}(t) + \alpha_{36} X_{\text{NAO}}(t) X_{\text{Wind}}(t). \\
X_{\text{Cope}}(t+1) &= \alpha_{37} X_{\text{LargeCope}}(t) + \alpha_{38} X_{\text{FishLarvae}}(t) + \alpha_{39} X_{\text{Wind}}(t) + \\
&\quad \alpha_{40} X_{\text{LargeCope}}(t) X_{\text{Wind}}(t).
\end{aligned}
$$

Appendix G

Bayesian state-space model

The Bayesian state-space model that I am proposing can be described as follows:

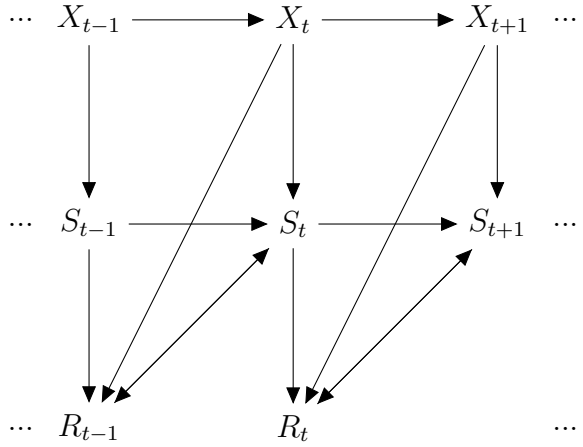
$$S_{t+1} = S_t \exp(-m - f) + R_t X_{t+1} + \epsilon_{t+1}, \tag{G.1}$$

$$X_{t+1} = \phi X_t + w_{t+1}, \tag{G.2}$$

where S_{t+1} is the stock size at time $(t+1)$, R_t is the recruits at time t. The parameters m and f are the natural and fishing mortality rates respectively. The variables ϵ_{t+1} and w_{t+1} are zero mean identically independently distributed Gaussian noise on the measurements and underlying process respectively. I describe the assessment noise as a Gaussian heteroscedastic noise such that $\epsilon_{t+1} \sim \mathcal{N}(0, e^{(\eta_0+\eta_1 S_{t+1})})$ where η_0 and η_1 are unknown parameters, and the underlying process noise as $w_{t+1} \sim \mathcal{N}(0, \sigma^2)$.
Equation (G.1) represents the observation density $S_{t+1} \sim p(S_{t+1}|X_{t+1})$ which is capable of generating one-step ahead forecast of the population size. Equation (G.2) represents the state evolution density $X_{t+1} \sim p(X_{t+1}|X_t)$ which is a first order Markov process. Figure G.1 illustrates the conditional independence structure in a graphical model in which directed arrows represent conditional dependencies in the set of conditional distributions comprising the model.

Beddington and May (1977) stated that the size of the population is maintained at an equilibrium point controlled by the variability of recruitment, which balances the losses resulted from natural mortality and low harvesting rates. This inspired me to integrate a latent process X_{t+1} to the observation density by multiplying it to the recruitment random variable R_t. I identify this latent process as the *recruitment variability process*. This process models both density dependent

Figure G.1: Directed graph describing the proposed AHMM in which the distribution of the observation S_t depends on previous observation S_{t-1}, on the latent state X_t, and on R_{t-1}.



and density independent factors. Environmental changes, food supplies and harvesting are density independent factors. In contrast, intraspecific competitions, predations and diseases are classified as density dependent limiting factors that regulate the recruitment variation in marine fishes in the early life history (Myers and Cadigan, 1993b). In normal situations, X_{t+1} would be close to one, but when the conditions become adverse its value would drop near to zero indicating a high variability in recruits. Intermediate values would indicate a smoothly varying profile. This description restricts the temporal variability of X_{t+1} to be defined between 0 and 1.

For convenience, I propose a first order Markov process for the state evolution density. The regulation process X_{t+1} is stationary if $-1 < \phi < 1$, hence the mean can be described as

$$\begin{aligned} & \mathbb{E}(X_{t+1}) = \phi\mathbb{E}(X_t) + \mathbb{E}(\tau_{t+1}) \\ \Rightarrow\ & \mathbb{E}(X_{t+1}) = \phi\mathbb{E}(X_{t+1}) + 0 \\ \Rightarrow\ & \mathbb{E}(X_{t+1}) = 0. \end{aligned}$$

and the variance as

$$\begin{aligned} \mathbb{V}(X_{t+1}) &= \phi^2 \mathbb{V}(X_t) + \mathbb{V}(\tau_{t+1}) \\ \Rightarrow \mathbb{V}(X_{t+1}) &= \phi^2 \mathbb{V}(X_{t+1}) + \sigma^2 \\ \Rightarrow \mathbb{V}(X_{t+1}) &= \frac{\sigma^2}{1-\phi^2}. \end{aligned}$$

I fix $\sigma^2 = 1 - \phi^2$ to set the marginal variance of the state evolution density at stationarity to 1, such that $X_{t+1} \sim \mathcal{N}(0, 1)$. This would simulate all the components from a normal distribution. Elements of the state evolution density $\mathbf{X} = [X_1, \ldots, X_{t+1}]$ represent the relative regulation strength at different time points.

As all parameters are static and do not change over time, the full joint density over all quantities over $t = 0 : T$ can be described as

$$\begin{aligned} p(\mathbf{S}, \mathbf{R}, m, f, \mathbf{X}) = &\prod_{t=1}^{T} p(S_{t+1} \mid S_t, R_t, m, f, X_{t+1}, \eta_0, \eta_1) p(X_1) \\ &\times \prod_{t=2}^{T} p(X_{t+1} | X_t, \phi, \sigma) \pi(\eta_0) \pi(\eta_1) \pi(\sigma) \pi(\phi). \end{aligned} \tag{G.3}$$

I use $\boldsymbol{\theta} = \{m, f, \eta_0, \eta_1, \phi\}$ to denote collectively the model parameters. Next, one may substitute the heteroscedastic noise with other possible noises, such that (1) an isotropic noise and (2) a heavy tailed non-Gaussian innovations at the observational level.

All parameters are considered as static, but one may propose a time varying parameters to accommodate with possible temporal heterogeneity. A change-point model has the role to reveal the temporal heterogeneity in a sequence of observations by estimating the number of change-points, n , and their position $\tau_1, \tau_2, \ldots, \tau_n$. It turns out that observations are homogeneous within segments and heterogeneous across segments. Each segment is governed by a set of parameters θ and independent of the parameters of other segments which implies that the change-points satisfy a Markov property.

Appendix H

Reprint of publication I

Panikian, G. and Cussens, J. and Pitchford, W.J. (2015): Identification and quantification of heteroscedasticity in stock-recruitment relationships. *Canadian Journal of Fisheries and Aquatic Sciences*, NRC Research Press, 72, pp 1259-1271, 2015.

Volume 72

2015

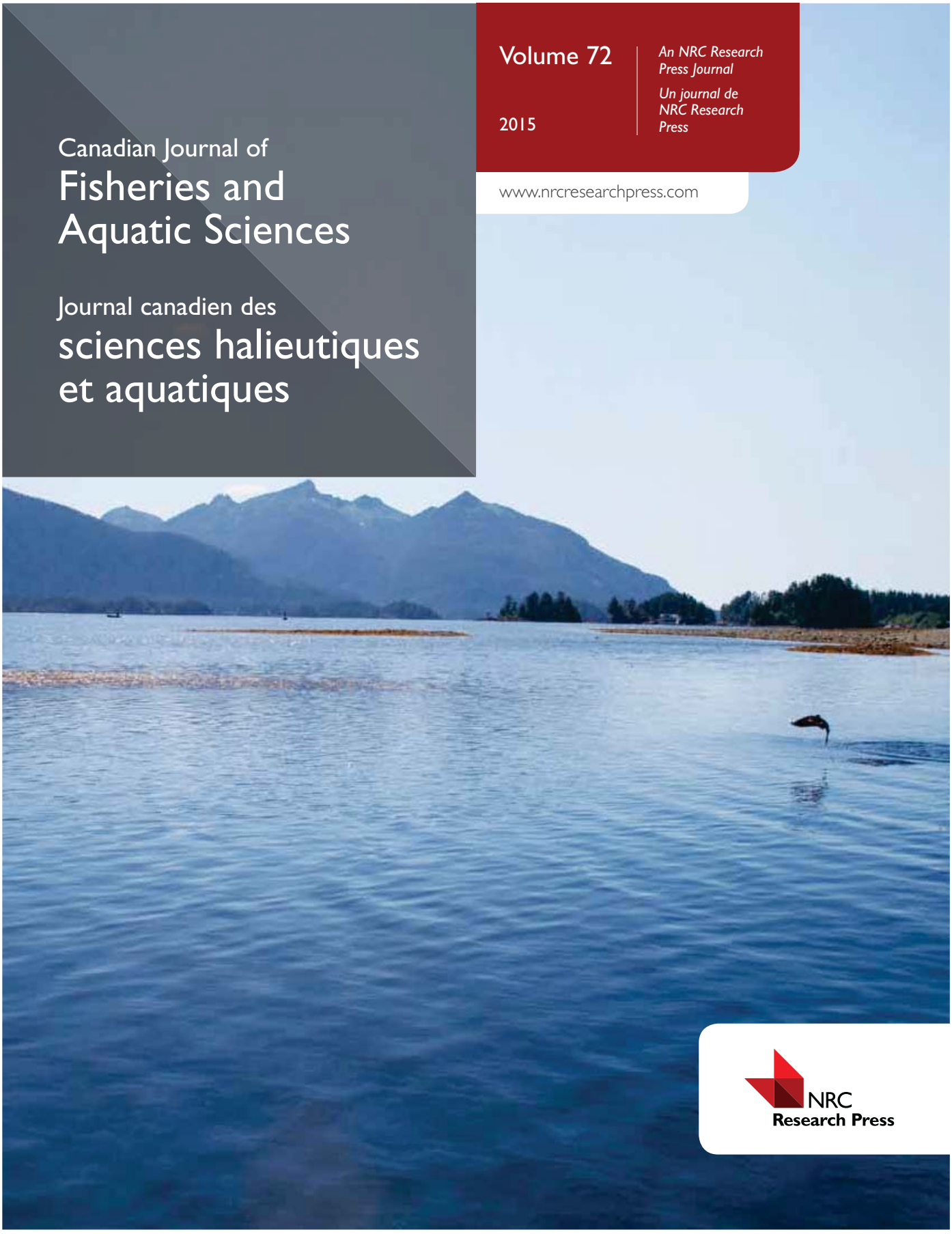
An NRC Research Press Journal

Un journal de NRC Research Press

www.nrcresearchpress.com

Canadian Journal of Fisheries and Aquatic Sciences

Journal canadien des sciences halieutiques et aquatiques



NRC Research Press

NRC Research Press

ARTICLE

Identification and quantification of heteroscedasticity in stock–recruitment relationships

Garo Panikian, James Cussens, and Jonathan W. Pitchford

Abstract: Nonconstant variance (heteroscedasticity) in the stock–recruitment (S-R) relationship is proposed as an important factor in sustainable fisheries management, but its reliable estimation from noisy populations is problematic. We developed methods for both frequentist and Bayesian approaches to test whether we can accurately estimate the degree of heteroscedasticity in 90 published S-R populations. We estimated the confidence interval for the heteroscedastic regression model via a parametric bootstrap approach and the credible interval for the Bayesian method via a Markov chain Monte Carlo sampling algorithm. We found strong evidence of negative heteroscedasticity in several stocks, regardless of the statistical paradigm, the details of density dependence, and the methods used to generate the original populations. This statistical framework, together with its associated freely available software, provides an efficient and reliable setting for assessing heteroscedasticity of the S-R relationship in fisheries.

Résumé : Si le caractère non constant de la variance (hétéroscédasticité) de la relation stock–recrutement (S-R) est proposé comme facteur important dans la gestion durable des pêches, l'estimation fiable de cette propriété pour des populations bruitées pose problème. Nous avons mis au point des méthodes pour des approches tant fréquentistes que bayésiennes pour vérifier s'il est possible d'estimer avec exactitude le degré d'hétéroscédasticité dans 90 populations de S-R publiées. Nous avons estimé l'intervalle de confiance pour le modèle de régression hétéroscédastique par une approche d'amorçage paramétrique et l'intervalle de crédibilité pour la méthode bayésienne par un algorithme d'échantillonnage de Monte Carlo par chaînes de Markov. Nous avons constaté de fortes indications d'une hétéroscédasticité négative dans plusieurs stocks, quels que soient le paradigme statistique, les détails de la dépendance sur la densité et les méthodes utilisées pour générer les populations initiales. Ce cadre statistique, combiné à un logiciel associé gratuit, offre un contexte efficient et fiable pour évaluer l'hétéroscédasticité de la relation S-R dans les pêches. [Traduit par la Rédaction]

Introduction

Reliable mathematical modelling and prediction of fish populations is of great importance socially and economically, as well as being a necessary ingredient in the conservation of biodiversity. Various natural and anthropogenic factors affect fish populations, with the life of juvenile fish typically being characterized by enormous mortality rates (Hilborn and Walters, 1992). Newly hatched fish larvae have low probability of reaching adulthood (Pitchford et al. 2005). Mortality is due to variability in food supply, migration, predation, starvation, poisonous pollutants, and fishing activities (Steele 1977), resulting in an unpredictable relationship between the adult population (stock) and the juveniles (recruitment) that will successfully survive to enter the adult population in the future. Understanding the stock–recruitment (S-R) relationship therefore requires careful statistical techniques forming a crucial ingredient in the sustainable management of these exploited natural resources. From the point of view of sustainable management, Shepherd et al. (1990) studied plausible regulatory processes for analysing fish populations and argued that increased variability at low stock sizes might prevent the collapse of stocks subject to high mortality rates, a theme echoed by Minto et al. (2008). Hsieh et al. (2006) presented the first empirical evidence that fishing could increase the survival variability (a proxy for recruitment variability) in an exploited population and advocated that increased variability of exploited populations favours a precautionary management approach.

Heteroscedastic models (i.e., statistical models using nonconstant variance) have gained much interest in recent years to explain the regulatory mechanisms in fish populations. Minto et al. (2008) developed a stochastic method applied to a meta-analysis of 147 fisheries populations to argue that survival variability is inversely proportional to stock size. Their model was inspired by Peterman (1981), who argued that random variation in marine survival rates tends to follow a log-normal distribution, but the novelty of their method was to incorporate a functional form of nonconstant recruitment variability over adult abundance. More recently, Burrow et al. (2013) investigated the feasibility of applying heteroscedastic models in practice, using two North Sea stocks as examples. They uncovered a weakness of using a heteroscedastic regression model by showing it to be statistically unreliable to fit the parameters based on small S-R populations (containing 40 or 50 data points), but made a mistake while defining the log-likelihood function (i.e., missing a square term and a factor of 0.5) and restricting their analysis to only two populations. The use of heteroscedastic models is controversial because previous research engaged in interpreting nonconstant variance

Received 10 December 2014. Accepted 13 April 2015.

G. Panikian.* York Centre for Complex Systems Analysis, University of York, York, YO10 5GE, UK.
J. Cussens. Department of Computer Science, The Hub, Deramore Lane, University of York, York, YO10 5GE, UK; York Centre for Complex Systems Analysis, University of York, York, YO10 5GE, UK.
J.W. Pitchford. Departments of Biology and Mathematics, University of York, York, YO10 5YW, UK; York Centre for Complex Systems Analysis, University of York, York, YO10 5GE, UK.
Corresponding author: Garo Panikian (e-mail: gp547@york.ac.uk).
*Present address: Department of Computer Science, The Hub, Deramore Lane, University of York, York, YO10 5GE, UK.

has failed to provide a clear-cut answer about its reliable estimation for fisheries management.

The aims of this study were (*i*) to develop frequentist and Bayesian methods for accurately identifying the nonconstant variance exhibited in a density-dependent model and (*ii*) to test the reliability of these methods on 90 S-R populations. Since none of the S-R populations are direct observations, we select populations estimated by virtual population analysis (VPA)-type assessments so as to ensure that the recruitment estimates are derived from the catch-at-age data, which is not dependent on the estimate of the spawning stock biomass (SSB). We analyse the edge effects at the beginning and end of the time series data to test whether VPA methods have an impact upon our results. We employ the two dominant approaches to inference, known as Bayesian and frequentist statistical methods, to determine whether one can reliably estimate the nonconstant variance. We conclude that within either the frequentist or Bayesian paradigm, the reliability of determining the existence of a negative nonconstant variance can only be assessed on a case-by-case basis.

Materials and methods

We prune S-R populations collated in the publicly available RAM legacy database version 1 (www.ramlegacy.org; Ricard et al. 2012) by restricting the analysis only to those estimated by VPA-type assessments. The SSB is measured in tonnes; however, the recruitment is measured in thousands of individuals. The 12 VPA-type assessment methods classified under this category are as follows: VPA, SPA, XSA, FLXSA, ADAPT, NFT-ADAPT, B-ADAPT, SXSA, SPA-ADAPT, NFT-ADAP, ISVPA, and hybrid. VPA, also known as cohort analysis, follows cohorts through their whole life, using catch-at-age data and natural mortality to back-calculate what recruitment had to be to support the catch (Hilborn and Walters 1992). In contrast, assessments based on integrated analyses and statistical catch-at-age assessments employ an underlying S-R relationship, so fitting an S-R curve to their time series is not appropriate. There were 100 S-R populations obtained with VPA-type assessment, but 10 populations had missing data or no data at all. Accordingly, we restricted our analysis on the remaining 90 fish populations, representing 32 species (see Appendix A, Table A1).

The model

To understand the relationship between SSB and recruitment, Minto et al. (2008) proposed the following model, such that

$$\ln\left(\frac{R_i}{S_i}\right) \sim \text{i.i.d. } \mathcal{N}(\mu_i, \sigma_i^2) \quad \text{where} \quad \mu_i = \ln(\alpha) + \frac{1}{\gamma}\ln(1 - \gamma\beta S_i) \quad \text{and} \quad \sigma_i^2 = \exp(\eta_0 + \eta_1 S_i) \tag{1}$$

where R_i and S_i are the estimated number of recruits and SSB for each observation i, respectively. This is a regression model that assumes the logarithm of the ratio R_i/S_i is an independent and identically distributed (i.i.d.) sample from a Gaussian model with nonconstant variance. In practice, none of the populations (i.e., SSB and recruitment) are actually direct observations. They are in reality model outputs (parameter estimates) from fisheries assessments, where models have been previously fitted to fisheries data (catch, age structure information, indices, etc.). The parameters α and β measure the productivity and the density-dependent mortality (capacity) in a population, respectively. The density-independent part of the variance is described by η_0, with the density-dependent variance described by η_1, known as the heteroscedastic coefficient. The parameter γ enables us to choose among several survival models. For instance, $\gamma = -1000$ generates a model with essentially no density dependence, and increasing γ increases the amount of density dependence, reproducing several models that have been advocated in previous studies (Minto et al. 2008), such as: $\gamma = -2$ (Cushing-like), $\gamma = -1$ (Beverton–Holt), $\gamma = 0$ (Ricker), and $\gamma = 1$ (Schaefer).

Likelihood of the model

We let $\boldsymbol{R} = (R_1, R_2, \ldots, R_n)$ and $\boldsymbol{S} = (S_1, S_2, \ldots, S_n)$ be the respective recruitment and stock model input vectors. The log-likelihood of the heteroscedastic regression model is

$$\mathcal{L}\{\ln(\boldsymbol{R}/\boldsymbol{S}), \alpha, \beta, \eta_0, \eta_1\} = -\frac{n}{2}\ln(2\pi) - \frac{1}{2}\sum_{i=1}^{n}(\eta_0 + \eta_1 S_i) - \frac{1}{2}\sum_{i=1}^{n}\frac{\left\{\ln\left(\frac{R_i}{S_i}\right) - f\left(\frac{R_i}{S_i}\right)\right\}^2}{\exp(\eta_0 + \eta_1 S_i)} \tag{2}$$

where

$$f\left(\frac{R_i}{S_i}\right) = \ln(\alpha) + \frac{1}{\gamma}\ln(1 - \gamma\beta S_i)$$

and n is the number of observations. We fix $\gamma = -1$ for our analyses here (representing the Beverton–Holt compensation model), which turns the log-likelihood function into the form

$$\mathcal{L}\{\ln(\boldsymbol{R}/\boldsymbol{S}), \alpha, \beta, \eta_0, \eta_1\} \propto -\frac{1}{2}\sum_{i=1}^{n}(\eta_0 + \eta_1 S_i) - \frac{1}{2}\sum_{i=1}^{n}\frac{\left\{\ln\left(\frac{R_i}{S_i}\right) - \ln(\alpha) + \ln(1 + \beta S_i)\right\}^2}{\exp(\eta_0 + \eta_1 S_i)} \tag{3}$$

For a constant variance, the coefficient of heteroscedasticity is zero and the variance would be written as $\sigma^2 = \exp(\eta_0)$. We choose to scale both SSB and recruitment model inputs with their maximum values respectively, as in (Minto et al. 2008).

To determine whether the log-likelihood function (eq. 3) is globally concave or not, we examine the matrix of second derivatives (or the Hessian matrix) from which we derive the sequence of determinants known as "principal minors". Since the leading principal minors do not alternate in sign, we conclude that the log-likelihood function is not concave, and hence we require suitable methods for solving the maximum likelihood problem.

Why choose a heteroscedastic regression model?

The Akaike information criterion (AIC) statistic (Akaike 1973) is a method used to select a model from a set of models; it penalizes the likelihood for the number of parameters that we estimate, such that

$$\text{AIC} = -2\mathcal{L}(\hat{\boldsymbol{\theta}}) + 2D \tag{4}$$

where $\mathcal{L}(\hat{\boldsymbol{\theta}})$ is the log-likelihood of the model evaluated at the maximum likelihood estimate of $\boldsymbol{\theta}$, and D is the number of (independent) model parameters. This AIC statistic indicates that a smaller value has a better fit of the model to the data. An alternative version of this statistic with a more severe penalty is known as the bias-corrected AIC, denoted by AIC_c (Hurvich and Tsai 1989), defined by

Table 1. Descriptive comparison of asymptotic and bootstrap methods for estimating the approximate 95% confidence interval for η_1.

			Asymptotic 95% CI		Bootstrap 95% CI	
Assessment ID	No. of samples	$\hat{\eta}_1$	Lower limit	Upper limit	Lower limit	Upper limit
AFWG-GHALNEAR-1960-2010	43	1.31	0.85	1.76	−1.24	3.16
AFWG-HADNEAR-1947-2010	58	−1.77	−2	−1.54	−3.42	0.31
AFWG-HADNS-IIIa-1963-2011	49	0.33	0.01	0.66	−2.4	2.72
AFWG-POLLNEAR-1957-2011	49	−0.98	−1.16	−0.79	−2.52	0.43
DFO-HAD5Zejm-1968-2003	34	1.56	1.12	1.99	−0.94	3.62
DFO-HERR4VWX-1964-2006	41	2.1	1.75	2.44	−0.27	4.05

$$\text{AIC}_c = -2\mathcal{L}(\hat{\boldsymbol{\theta}}) + 2D\frac{n}{n - D - 1} \tag{5}$$

This criterion avoids overfitting by replacing the penalty term of AIC with an exact expression for the bias adjustment and provides improved model selection for small samples. However, as n gets large, AIC_c converges to AIC, rendering the AIC_c a more effective statistic in practice. We compared model fitting for heteroscedastic and nonheteroscedastic regression models using the AIC_c statistic. We found a prevalence of the heteroscedastic model for 78 out of 90 populations, showing that the heteroscedastic model had a much better fit across the majority of stocks, regardless of the coefficient of heteroscedasticity. Since the sign of η_1 has a great influence in determining whether such a model is appropriate for devising optimal harvest strategies, we investigated whether we can reliably estimate the sign of η_1; we also applied the AIC_c statistic as a measure to determine the most appropriate value for γ that fits the S-R populations.

Frequentist inference

Let $\hat{\boldsymbol{\theta}}$ be the maximum likelihood estimation (MLE) of $\boldsymbol{\theta} = (\alpha, \beta, \eta_0, \eta_1)$ that maximizes the log-likelihood function $\mathcal{L}(\boldsymbol{R}|\boldsymbol{\theta})$. Here, we make use of the software AD Model Builder (ADMB) to solve this optimization problem (Fournier et al. 2012) because we found it to perform better than off-the-shelf Nelder-Mead and Newton methods.

To assess the estimation error for η_1 (heteroscedasticity parameter), we first employed ADMB to get standard errors and confidence intervals using the standard likelihood asymptotics, but we were confronted with two hurdles. First, we found 42 out of 90 populations had a sample size of fewer than 30 data points (among them two populations found with Hessian matrix not positive definite) for which it is common not to apply the Central Limit Theorem. Second, among the remaining 48 populations, we found a case (namely NWWG-HADFAPL-1955-2011) for which the Hessian is not positive definite, rendering the confidence intervals unobtainable. Hence, we decided to abandon the standard likelihood asymptotics method and use the parametric bootstrap sampling approach, as in DiCiccio and Efron (1996). This method is also known as bootstrapping raw data, where each replication is obtained by sampling from the heteroscedastic distribution fitted with the MLE $\hat{\boldsymbol{\theta}}$. The theory of this method shows that the bootstrap confidence intervals are second-order correct as well as second-order accurate (DiCiccio and Efron 1996, sections 8 and 9), and it is appropriate for studies with small sample size. We describe this sampling method for simulating new recruits as follows:

Step 1: Use both stock and recruitment model inputs to estimate $\hat{\boldsymbol{\theta}} = (\hat{\alpha}, \hat{\beta}, \hat{\eta}_0, \hat{\eta}_1)$, the MLE of $\boldsymbol{\theta} = (\alpha, \beta, \eta_0, \eta_1)$.

Step 2: Draw i.i.d. samples $\boldsymbol{x}^* = (x_1^*, x_2^*, \ldots, x_n^*)$ from $\mathcal{N}(\hat{\mu}_i, \hat{\sigma}_i^2)$ where

$$\hat{\mu}_i = \ln(\hat{\alpha}) - \ln(1 + \hat{\beta}S_i) \quad \text{and} \quad \hat{\sigma}_i^2 = \exp(\hat{\eta}_0 + \hat{\eta}_1 S_i)$$

with n being the number of data points found in the S-R population, and $i = 1, \ldots, n$.

Step 3: Simulate new recruit $\boldsymbol{R}^* = \boldsymbol{S}\exp(\boldsymbol{x}^*)$, such that $\boldsymbol{R}^* = (R_1^*, R_2^*, \ldots, R_n^*)$.

Step 4: Scale $\boldsymbol{R}^*$ with its maximum value.

Step 5: Refit the regression model to the simulated data $(\boldsymbol{R}^*, \boldsymbol{S})$ and estimate $\hat{\boldsymbol{\theta}}^*(\boldsymbol{R}^*|\boldsymbol{S})$.

Step 6: Repeat steps 2 to 5, 1000 times, to obtain a good approximation of the confidence interval.

We compared the parametric bootstrap method with the asymptotic confidence intervals from ADMB for the subset of populations with adequate sample sizes (greater than 30) and positive definite Hessians; we found that the bootstrap method provides empirical coverages for $\hat{\eta}_1$ noticeably wider than the asymptotic confidence intervals (Table 1). The findings confirm that under the first-order asymptotic theory, the residual errors of recruitments are not normally distributed, rendering the first-order asymptotic theory inappropriate to assess the estimation error for η_1. To account for possible skewness of the estimator, Singh (1981) and DiCiccio and Efron (1996) proved that second-order properties are often more desirable, as they improve by an order of magnitude upon the accuracy of the standard intervals.

During this analysis we found the Hessian matrix for three populations, namely INIDEP-PATGRENADIERSARG-1983-2006, NWWG-HADFAPL-1955-2011, and NWWG-HERRIsum-1984-2011 (also known by their ID numbers: 25, 58, and 60), not positive definite, meaning that the optimizer will fail to find the true maximum of the log-likelihood function for which the parametric bootstrapping method would result in estimating incorrect MLEs. To overcome this hurdle, one can apply a Bayesian approach to estimate the posterior distribution of η_1, as discussed in the section on Bayesian inference.

Our above analysis could be incomplete, because we focused only on the Beverton–Holt compensation model (see eq. 3). Here we generalize our previous assumption by making available the set of possible models $\gamma \in \{-2, -1, 0, +1\}$ and choose the one that provides the minimum AIC_c score for each population, respectively. A nonasymptotic recruitment is obtained for $\gamma = -2$, which means that recruitments can grow with adult abundance size. However, an overcompensation model is obtained for $\gamma \in \{-1, 0, +1\}$, which are different sorts of density dependence models. For $\gamma \in \{-2, -1, 0, +1\}$, we obtain different models but with the same number of parameters $\{\alpha, \beta, \eta_0, \eta_1\}$. We describe the log-likelihood functions of these models as follows: the Cushing-like model is obtained for $\gamma = -2$:

$$\mathcal{L}\{\ln(\boldsymbol{R}/\boldsymbol{S}), \alpha, \beta, \eta_0, \eta_1\} \propto -\frac{1}{2}\sum_{i=1}^{n}(\eta_0 + \eta_1 S_i) - \frac{1}{2}\sum_{i=1}^{n}\frac{\left\{\ln\left(\frac{R_i}{S_i}\right) - \ln(\alpha) + \frac{1}{2}\ln(1 + 2\beta S_i)\right\}^2}{\exp(\eta_0 + \eta_1 S_i)} \tag{6}$$

The Beverton–Holt model is obtained for $\gamma = -1$:

$$\mathcal{L}\{\ln(\boldsymbol{R}/\boldsymbol{S}), \alpha, \beta, \eta_0, \eta_1\} \propto -\frac{1}{2}\sum_{i=1}^{n}(\eta_0 + \eta_1 S_i) - \frac{1}{2}\sum_{i=1}^{n}\frac{\left\{\ln\left(\frac{R_i}{S_i}\right) - \ln(\alpha) + \ln(1 + \beta S_i)\right\}^2}{\exp(\eta_0 + \eta_1 S_i)} \quad (7)$$

The Ricker model is obtained after developing a first-order Taylor expansion of the $\log(1 - \gamma\beta S_i)$ function around $\gamma \to 0$, resulting in $-\beta S_i$ after dividing it by γ, which takes the form

$$\mathcal{L}\{\ln(\boldsymbol{R}/\boldsymbol{S}), \alpha, \beta, \eta_0, \eta_1\} \propto -\frac{1}{2}\sum_{i=1}^{n}(\eta_0 + \eta_1 S_i) - \frac{1}{2}\sum_{i=1}^{n}\frac{\left\{\ln\left(\frac{R_i}{S_i}\right) - \ln(\alpha) + \beta S_i\right\}^2}{\exp(\eta_0 + \eta_1 S_i)} \quad (8)$$

The Schaefer model is obtained for $\gamma = +1$:

$$\mathcal{L}\{\ln(\boldsymbol{R}/\boldsymbol{S}), \alpha, \beta, \eta_0, \eta_1\} \propto -\frac{1}{2}\sum_{i=1}^{n}(\eta_0 + \eta_1 S_i) - \frac{1}{2}\sum_{i=1}^{n}\frac{\left\{\ln\left(\frac{R_i}{S_i}\right) - \ln(\alpha) - \ln(1 - \beta S_i)\right\}^2}{\exp(\eta_0 + \eta_1 S_i)} \quad (9)$$

We use the AIC_c score with the point estimate approach for applying model selection instead of approximating the marginal likelihood function. We applied ADMB over the four different models (e.g., $\gamma = \{-2, -1, 0, +1\}$), and we selected for each population the model with the minimum AIC_c score. To find a point estimate for $\gamma = +1$, the parameter β should be less than $\{1/\max(S_i)\} = 1$ so as to assert a valid argument for the logarithmic function — both $\boldsymbol{S}$ and $\boldsymbol{R}$ are respectively scaled with their maximum value. The purpose of this analysis is to provide a precise assessment for the nonconstant variance (instead of relying only on the Beverton–Holt model) as it fits the data more accurately; we found a prevalence of the Schaefer model for the majority of populations that underscores the importance of using density-dependent models in explaining the S-R relationships (Table 2). To better understand and evaluate the impact of the heteroscedasticity coefficient, we illustrate in Fig. 1 plots showing recruits versus relative spawning stock biomass along with the estimated stock–recruit relationship and approximate 95% confidence intervals around this relationship for both heteroscedastic and nonheteroscedastic (i.e., constant variance) models; these plots are for four populations fitted with different shape parameter γ. Recruitment is commonly assumed to have stochastic variability that follows a log-normal distribution from which we derive the expected recruits for each shape parameter γ, such as

$$\mathbb{E}(R) = \frac{\alpha \boldsymbol{S}}{\sqrt{1 + 2\beta \boldsymbol{S}}}\exp\left\{\frac{\exp(\eta_0 + \eta_1 \boldsymbol{S})}{2}\right\} \quad \text{for} \quad \gamma = -2 \quad (10)$$

$$\mathbb{E}(R) = \frac{\alpha \boldsymbol{S}}{1 + \beta \boldsymbol{S}}\exp\left\{\frac{\exp(\eta_0 + \eta_1 \boldsymbol{S})}{2}\right\} \quad \text{for} \quad \gamma = -1 \quad (11)$$

Table 2. Populations fitted with best-fit model parameter γ.

γ	Model	Fitted populations
–2	Cushing	12 out of 90
–1	Beverton–Holt	26 out of 90
0	Ricker	19 out of 90
+1	Schaefer	33 out of 90

Note: The model selection is based on the shape parameter γ corresponding to $\gamma = -2$ (Cushing-like), $\gamma = -1$ (Beverton–Holt), $\gamma = 0$ (Ricker), and $\gamma = 1$ (Schaefer).

$$\mathbb{E}(R) = \alpha \boldsymbol{S}\exp\left\{-\beta \boldsymbol{S} + \frac{\exp(\eta_0 + \eta_1 \boldsymbol{S})}{2}\right\} \quad \text{for} \quad \gamma = 0 \quad (12)$$

$$\mathbb{E}(R) = \alpha \boldsymbol{S}(1 - \beta \boldsymbol{S})\exp\left\{\frac{\exp(\eta_0 + \eta_1 \boldsymbol{S})}{2}\right\} \quad \text{for} \quad \gamma = 1 \quad (13)$$

The expected S-R curve is plotted based on the fitted value of γ, the MLE found with ADMB, and the SSB model input. We illustrate the heteroscedastic expected recruitment curve with solid black plot, and the nonheteroscedastic expected recruitment curve with dotted black plot — obtained by setting $\eta_1 = 0$. We construct the approximate 95% confidence interval for recruitment as follows: sort the SSB population in an ascending order; use eq. 1 to generate 10 000 samples for each element; approximate the 95% confidence interval of recruitment estimates for each SSB data point, using the percentile of the sampling distribution. Figure 1 shows that the coefficient of heteroscedasticity η_1 has a positive impact in estimating the approximate 95% confidence interval; the coverage of recruits (dashed plot) is more accurate than the nonheteroscedastic model (grey area) and hence its importance in fisheries management. The plots for the 90 S-R populations are illustrated in the Supplementary material[1] section.

Measure of confidence interval

We use the bias-corrected and accelerated method (BC_a; DiCiccio and Efron 1996) to form the approximate 95% confidence interval of the density distribution $\hat{\boldsymbol{\eta}}_1^*$. Let $\hat{G}(\hat{\eta}_1)$ be the cumulative distribution function of bootstrap replications $\hat{\boldsymbol{\eta}}_1^*$:

$$\hat{G}(\hat{\eta}_1) = \#\left(\hat{\boldsymbol{\eta}}_1^* \le \hat{\eta}_1\right)/B \quad (14)$$

In this case, $B = 1000$ replications. By definition, the bias-corrected $\kappa/2$ endpoints for the percentile bootstrap confidence interval are calculated as

$$\hat{\eta}_{1_{BC_a}}(\kappa) = \hat{G}^{-1}\left\{\Phi\left(\hat{z}_0 + \frac{\hat{z}_0 + z^{(\kappa)}}{1 - a\left(\hat{z}_0 + z^{(\kappa)}\right)}\right)\right\} \quad (15)$$

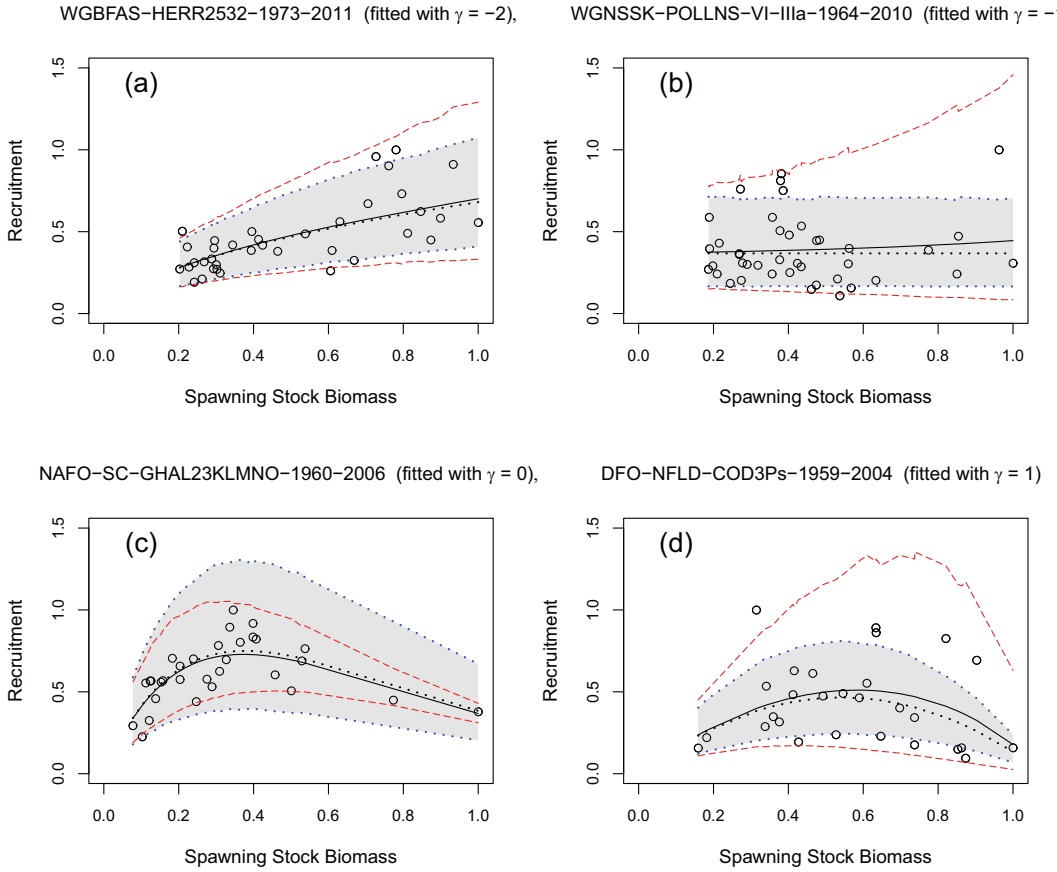
where $\Phi(.)$ is the standard normal cumulative distribution function. The BC_a interval is controlled by two parameters, namely the bias-correction $\hat{z}_0$ and acceleration parameters a. The bias-correction estimate $\hat{z}_0$ gives the proportion of estimates $\hat{\boldsymbol{\eta}}_1^*$ less than $\hat{\eta}_1$, such that

$$\hat{z}_0 = \Phi^{-1}\left\{\hat{G}(\hat{\eta}_1)\right\} = \Phi^{-1}\left\{\frac{\#\left(\hat{\boldsymbol{\eta}}_1^* \le \hat{\eta}_1\right)}{1000}\right\} \quad (16)$$

However, the acceleration parameter a measures how rapidly standard error changes on a normalized scale, which has an

[1]Supplementary data are available with the article through the journal Web site at http://nrcresearchpress.com/doi/suppl/10.1139/cjfas-2014-0549. This data show the expected stock–recruitment curves with approximate 95% confidence intervals fitted with different values of γ, for each of the 90 stocks analysed.

Fig. 1. Expected stock–recruitment curves with approximate 95% confidence intervals fitted with different values of γ. Examples of the herring, pollock, Greenland halibut, and cod families, chosen to illustrate the difference in fit between the heteroscedastic and nonheteroscedastic models, are shown: (*a*) herring from Eastern Baltic (fitted with $\gamma = -2$), (*b*) pollock from IIIa, VI, and North Sea (fitted with $\gamma = -1$), (*c*) Greenland halibut from Labrador Shelf – Grand Banks (fitted with $\gamma = 0$), and (*d*) cod from St. Pierre Bank (fitted with $\gamma = +1$). The expected recruit for the nonheteroscedastic model (dotted black plot) and its approximate 95% confidence interval (grey area) are compared against the expected recruit for the heteroscedastic model (solid black plot) and its approximated 95% confidence interval (dashed plot).



interpretation of skewness. A nonparametric estimate of a can be described as

$$\hat{a} = \frac{1}{6}\frac{\sum_{i=1}^{n} U_i^3}{\left(\sum_{i=1}^{n} U_i^2\right)^{3/2}} \quad (17)$$

We use the jackknife influence function to calculate U_i, where

$$U_i = (n-1)(\hat{\eta}_1 - \hat{\eta}_{1(i)}) \quad (18)$$

where $\hat{\eta}_{1(i)}$ is the estimate of η_1 based on the reduced data $\boldsymbol{R}_{(i)} = (R_1, \ldots, R_{i-1}, \ldots, R_n)$ and $\boldsymbol{S}_{(i)} = (S_1, \ldots, S_{i-1}, S_{i+1}, \ldots, S_n)$. Therefore, the central 95% BC_a interval for η_1 is given by

$$CI_{95\%}(\eta_1) = (\hat{\eta}_{1_{BC_a}}(0.025), \hat{\eta}_{1_{BC_a}}(0.975)) \quad (19)$$

Note that the confidence interval of η_1 is mainly influenced by the number of data points found in the population. The more data points we have, the narrower confidence interval we obtain. This means that the variability for $\hat{\eta}_1$ is small for large populations, thereby leading to a more reliable fit of the heteroscedastic model than for small population sizes. Figure 2 illustrates the approximate 95% BC_a confidence interval width versus the population size, applied to all populations.

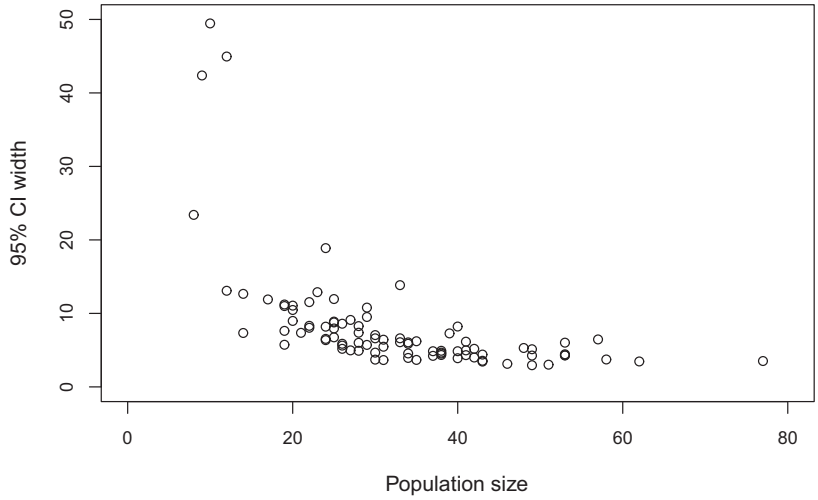
Classification based on the frequentist paradigm

Our goal is to investigate whether we could recover accurately the sign of the coefficient of heteroscedasticity. Here, we analyse whether the approximate confidence interval for η_1, denoted as $CI(\eta_1)$, falls in a region showing a consistent sign with the coefficient η_1. Accordingly, we classify each population in one of three ways: (–1): strong evidence for negative η_1 that is attained when $CI(\eta_1)$ falls in the negative region; (+1): strong evidence for positive η_1 when $CI(\eta_1)$ falls in the positive region; and (0): inconclusive evidence for heteroscedasticity.

Bayesian inference

Bayesian methods offer an alternative to the traditional frequentist method and may be particularly valuable for populations where there is already some information about the model's parameters. To define a Bayesian analogue of the method of fitting parameters outlined above, we need to specify prior distributions for α, β, η_0, and η_1 to quantify our knowledge before considering the data. Because we do not have prior knowledge for the param-

Fig. 2. Plot showing the effect of the sample size on the width of the approximated 95% confidence interval. This plot is generated for a Beverton–Holt stock–recruitment model ($\gamma = -1$).



eter values, we chose two arbitrary sets of priors to see to what extent the marginal posteriors vary. Firstly, we choose to define a normal prior information for $\log(\alpha)$, a wide uniform prior for β covering a region of positive values (to avoid numerical failures), and a symmetrical uniform prior around the origin for both η_0 and η_1, such that

$$\pi_1\{\log(\alpha)\} = \mathcal{N}(1, 5^2) \quad (20)$$

$$\pi_1(\beta) = \mathcal{U}(0, 6000) \quad (21)$$

$$\pi_1(\eta_0) = \mathcal{U}(-30, +30) \quad (22)$$

$$\pi_1(\eta_1) = \mathcal{U}(-30, +30) \quad (23)$$

Secondly, we define a normal prior information for $\log(\alpha)$, a Gamma distribution for β to constrain it to positive values, and a Gaussian distribution for both η_0 and η_1, such that

$$\pi_2\{\log(\alpha)\} = \mathcal{N}(1, 5^2) \quad (24)$$

$$\pi_2(\beta) = \text{Gamma}(1, 0.001) \quad (25)$$

$$\pi_2(\eta_0) = \mathcal{N}(0, 10^2) \quad (26)$$

$$\pi_2(\eta_1) = \mathcal{N}(0, 10^2) \quad (27)$$

It is common to assume independent priors for the parameters, such that $\pi(\boldsymbol{\theta}) = \pi\{\log(\alpha)\} \times \pi(\beta) \times \pi(\eta_0) \times \pi(\eta_1)$. We use Markov chain Monte Carlo (MCMC) sampling by JAGS (Just Another Gibbs Sampling; Plummer, 2003) from R via package rjags to sample from the joint posterior distribution $p(\alpha, \beta, \eta_0, \eta_1 \mid \boldsymbol{R}, \boldsymbol{S})$ with the purpose of estimating the marginal posterior distribution of η_1 given data.

Convergence criteria

We monitor the approximate convergence of MCMC by using the $\sqrt{\hat{R}}$ statistic provided in the Coda package in R. Gelman et al. (2004) described this statistic as a measure that compares variation between and within simulated sequences until "within" variation roughly equals "between" variation, for multiple parallel chains. One can be reasonably confident that convergence is achieved if $\sqrt{\hat{R}} < 1.1$. We simulated four parallel MCMC sequences of 10 000 iterations each after discarding 50 000 samples of each chain, referred to as "burn-in"; these chains are started each from a different initial value and thinned by taking one sample from every four samples so as to minimize the autocorrelation between samples. After convergence, each simulated sequence is close to the distribution of all other sequences combined together, which all converge to the same posterior distribution. If approximate convergence has not been reached, we identify those populations and repeat the approximation by increasing the number of burn-in samples (from 1e + 5 to 2e + 6) and even the number of step size adjustments (from 1e + 4 to 1e + 5) — tuned by n.adapt parameter. Figure 3 illustrates overlaid plots for the marginal posterior distribution of each parameter of interest ($\log(\alpha)$, β, η_0, η_1) with priors π_1 and π_2, respectively.

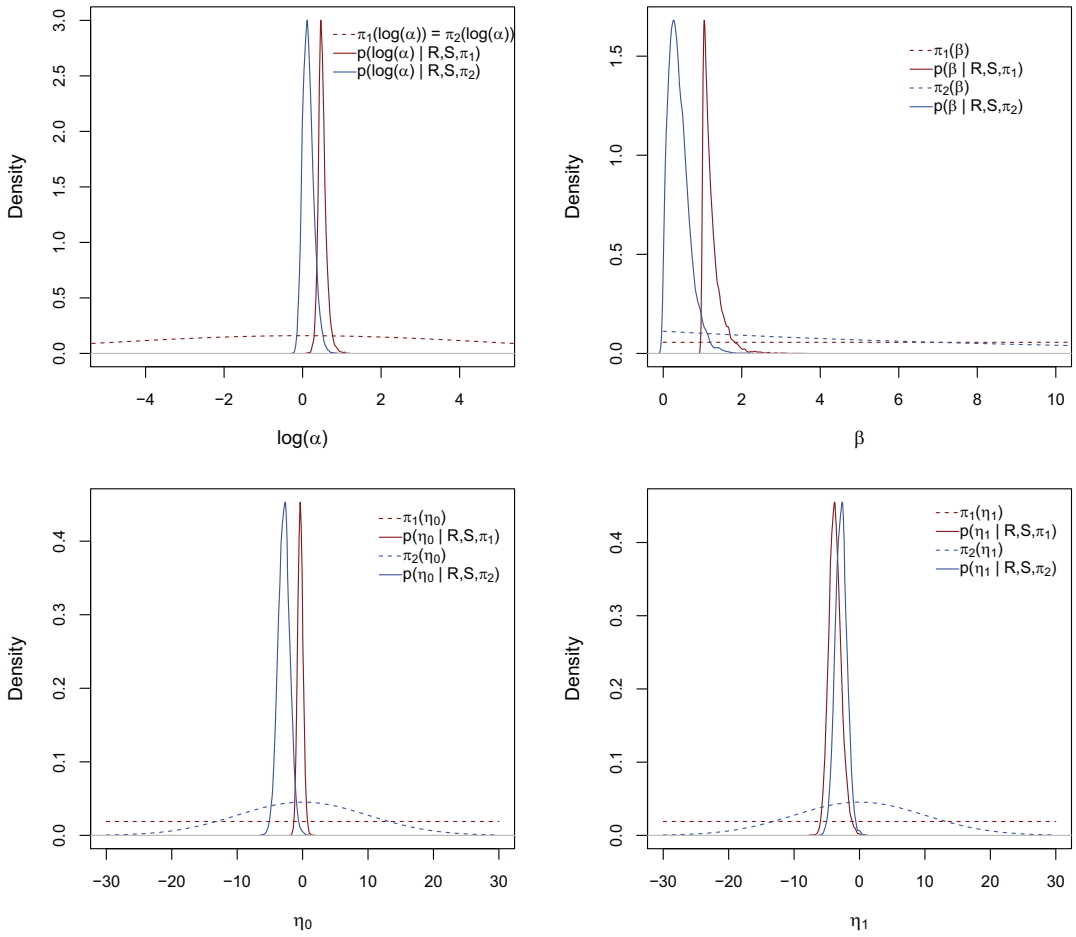
Bayesian sensitivity analysis

The theory of subjective probability enables one to apply a prior distribution in inference to reflect whatever is reasonably assumed. In Bayesian statistics, prior robustness is a real issue for inference; to reduce this concern, one should investigate whether slight changes in the prior distribution cause significant changes in the decision rule. Here we found that the choice between π_1 and π_2 did not influence the resulting Bayesian inference, indicating a reasonable degree of robustness.

Edge effects analysis

VPA estimates stock sizes and fishing mortality rates for each year class (cohort) making up the overall population; the recruitment is estimated as abundance at age 1, and the SSB is estimated by summing up stock sizes of age 2+ in each respective year (Anderson 1978). As we go backward in time, the final age class assumptions and the catch-at-age data totally drive the estimates to become very precise at the beginning of the age group; however, techniques based on the shrinkage to the mean factor — such as XSA — can impose constraints on the last year estimates as well as on the oldest age group (Daskalov 1998). To account for this kind of estimation error, we analyse the possibility of edge effects in the VPA methods by removing data points from the beginning and end of the time series data. Here we analyse two types of populations. First, we revisit results obtained from model selection by selecting populations having their approximate 95% confidence interval for η_1 falling entirely in the negative region (as

Fig. 3. Marginal posterior distributions for the parameters produced by the JAGS sampler sampled from priors π_1 and π_2. Each panel includes four density plots (except for the top left panel): two priors (π_1 and π_2) for each parameter and the posteriors corresponding to each of these priors when applied to the DFO-QUE-COD3Pn4RS-1964-2007 population.



shown in Appendix A, Table A1). Second, we select populations having more than 55 data points so as to check the effect on long time series data. The former selection presents seven full populations for which five were fitted with $\gamma = 1$, one was fitted with $\gamma = 0$, and another was fitted with $\gamma = -2$; the latter presents four full populations. Next we truncated two data points at both ends of the seven populations and five data points at both ends of the four biggest populations on which we repeated the analysis of testing the reliability of the nonconstant variance respectively. If the results were approximately matching to the pattern typically obtained with the full datasets, then we obtain evidence that edge effects in VPA methods are unlikely to influence our results; otherwise, we conclude that VPA methods are likely to influence the results.

Results

Statistical analysis based on the frequentist paradigm shows for a Beverton–Holt model the existence of seven populations having their approximate 99% confidence interval for η_1 falling entirely in the negative region (Table 3). Those seven populations are from six different fish species in six locations, indicating that this classification result is not peculiar to a particular species or location. Standard confidence levels were increased gradually to reflect the sensitivity of the classification labels of the 90 S-R populations to

Table 3. Confidence levels and data classification of the 90 S-R populations for a Beverton–Holt stock–recruitment model.

	Coding of $\hat{\eta}_1$ distribution		
Confidence level (%)	−1	0	+1
60	30	43	17
70	26	50	14
80	22	61	7
90	15	72	3
95	11	78	1
99	7	82	1

Note: The {−1, 0, +1} coding based on the $\hat{\eta}_1$ distribution indicates the presence of strong evidence for reliably identifying $\eta_1 < 0$, inconclusive evidence where the sign of η_1 cannot be identified, and a strong evidence for identifying $\eta_1 > 0$, respectively.

Fig. 4. Density plots of (*i*) 1000 parametric bootstrap replications of $\hat{\eta}_1$ (solid plot); (*ii*) marginal posterior distribution of η_1 with respect to π_1 (dot-dashed plot); and (*iii*) marginal posterior distribution of η_1 with respect to π_2 (dashed plot). The analysis is applied to the DFO-QUE-COD3Pn4RS-1964-2007 population.

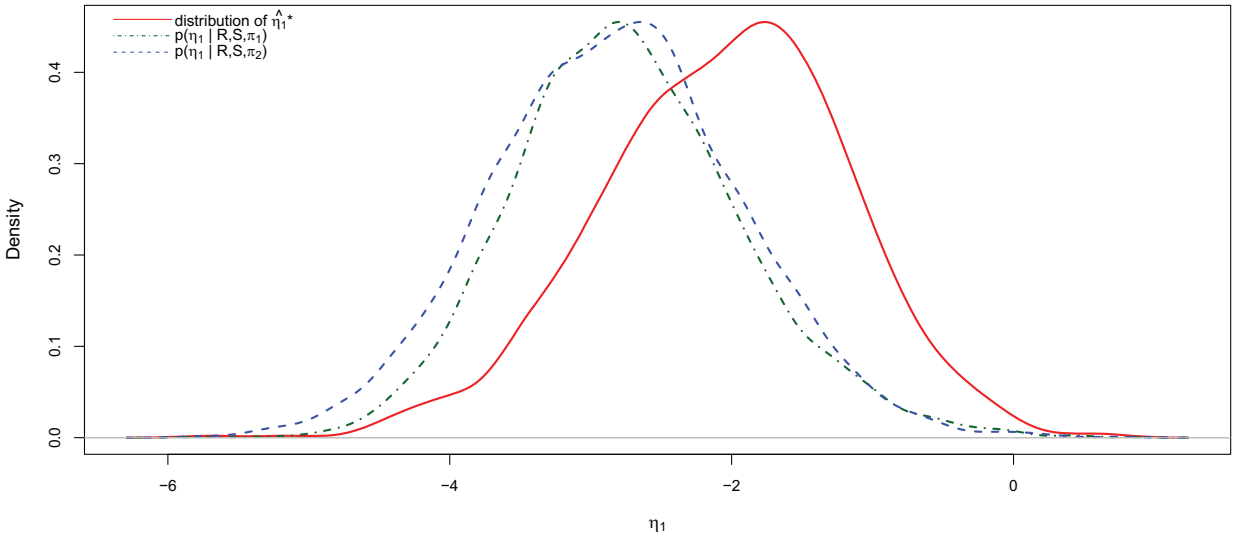


Table 4. Comparison between frequentist and Bayesian methods (with π_1 and π_2 priors) for a Beverton–Holt stock–recruitment model for evaluating the reliability of η_1 in survival across the 90 S-R fish populations.

	Coding of η_1 distribution								
	Frequentist			Bayesian π_1			Bayesian π_2		
Confidence level (%)	−1	0	+1	−1	0	+1	−1	0	+1
60	30	43	17	32	43	15	31	47	12
70	26	50	14	29	47	14	27	51	12
80	22	61	7	21	62	7	20	65	5
90	15	72	3	12	73	5	10	77	3
95	11	78	1	11	75	4	10	77	3
99	7	82	1	9	79	2	7	82	1

Table 5. Confidence levels and data classification of the 90 S-R populations using model selection.

	Label		
Confidence level (%)	−1	0	+1
60	31	48	11
70	28	53	9
80	23	62	5
90	16	70	4
95	7	82	1
99	3	87	0

Note: The {−1, 0, +1} coding based on the $\hat{\eta}_1$ distribution indicate the presence of strong evidence for reliably identifying $\eta_1 < 0$, inconclusive evidence where the sign of η_1 cannot be identified, and a strong evidence for identifying $\eta_1 > 0$, respectively.

the choice of cut-offs. For low confidence levels, we observe many populations classified with label −1, but this classification declines as the confidence level increases.

Next, we compared results obtained from the frequentist approach with those obtained from Bayesian methods. In the Bayesian framework, the credible interval is obtained from the marginal posterior distribution using the equal-tailed credible interval. Figure 4 illustrates a comparison between the frequentist and Bayesian inference (for different priors for η_1) applied to a single population, namely DFO-QUE-COD3Pn4RS-1964-2007. Note that in the frequentist method we are using a parametric bootstrap replication of $\hat{\eta}_1$; however, in the Bayesian setting we are estimating the marginal posterior distribution of η_1 given a particular population and a prior. We are in general interested in the outcome of these methods in knowing whether they produce the same result or not; the density of MLE simulates (solid line), the posterior distribution with respect to π_1 (dot-dashed line), and the posterior with respect to π_2 (dashed line) produced approximately comparable results. Their approximate 95% confidence interval and approximate 95% credible intervals were more likely to agree. We should also inform the reader that in this figure we used half of the posterior samples (i.e., 5000 samples) from both posteriors to avoid overlap in plots.

We generalized this comparison — by analysing the output of frequentist and Bayesian methods — for all 90 S-R populations, as illustrated in Fig. 5. The MLE used for bootstrap simulations is represented by an asterisk (*); the error bars on the asterisks represent the approximate 95% confidence intervals, and the square shape denotes the mode of simulated MLEs distribution. The error bars on the triangles represent the approximate 95% credible intervals with respect to π_1, and the error bars on the circles represent the approximate 95% credible intervals with respect to π_2. We observed a large approximate 95% confidence interval for the following population numbers: 9, 10, 13, 22, 25, 50, 55, 62, and 63, caused essentially by the small sample sizes of 12, 28, 29, 17, 20, 12, 8, 10, and 9 data points, respectively. In contrast, we found for some other populations (20, 50, 53, and 63) different marginal posteriors with respect to the choice of the prior. For the remaining populations we found robust posterior inference with respect to the choice of the prior (i.e., π_1 or π_2). Table 4 illustrates a comparison of the estimation error for η_1 assessed by the frequentist and Bayesian approaches when applied to the 90 S-R populations. We observed that both frequentist and Bayesian methods

Fig. 5. Comparison between the frequentist and Bayesian methods to inference for a Beverton–Holt model. The error bars on the asterisks show an approximate 95% BC_a confidence interval, where the asterisk symbol represents the MLE of η_1 and the square symbol represents the mode of simulated MLEs with bootstrapping. The error bars on the triangles and circles show the approximate 95% credible interval with respect to π_1 and π_2, respectively. The vertical axis represents the η_1 parameter, and the horizontal axis represents the sequential population number, ranging from 1 to 30, 31 to 60, and 61 to 90, respectively.

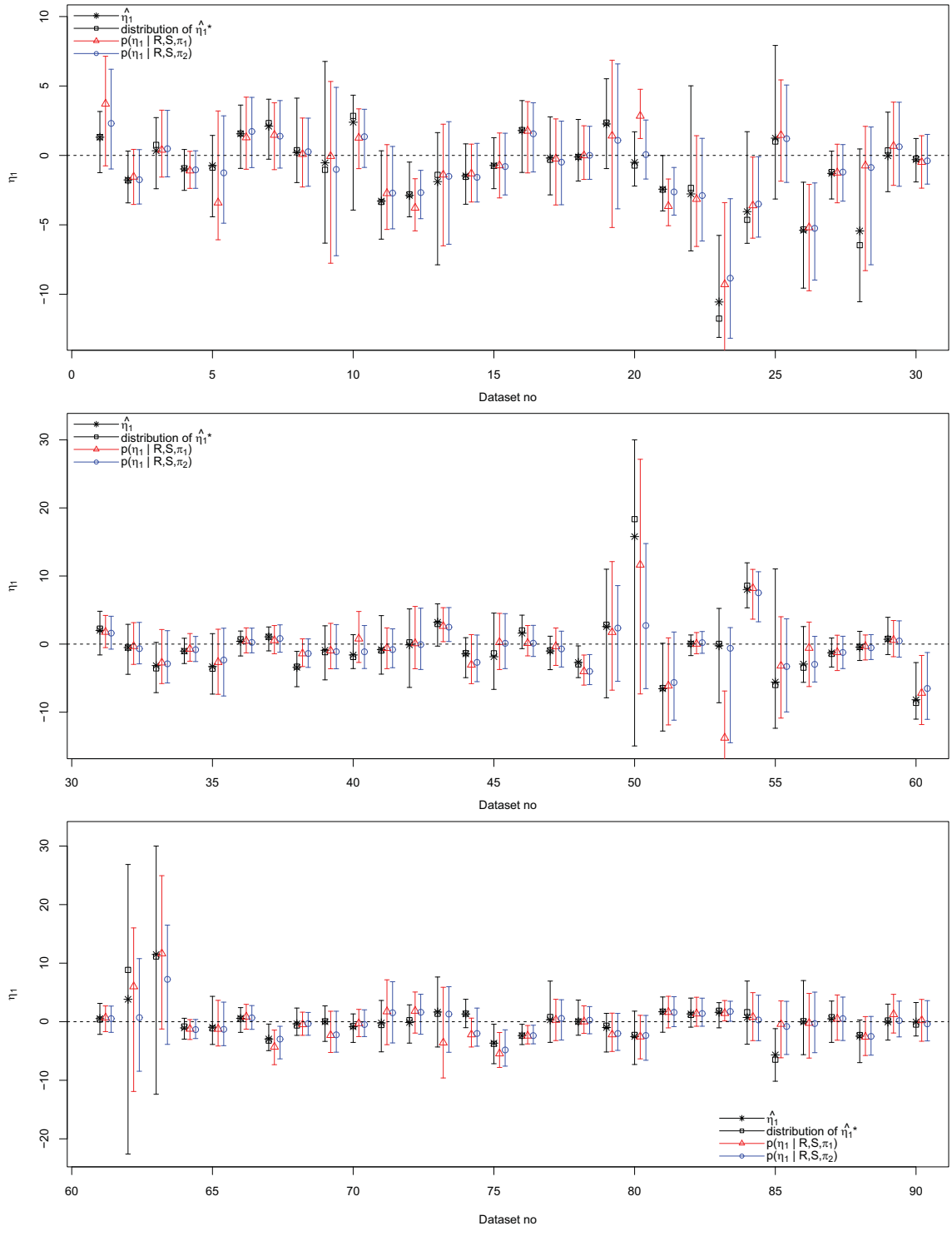


Table 6. Edge effect analysis applied to populations showing their approximate 95% confidence interval (CI) of η_1 falling in the negative region.

Assessment ID	γ	95% CI		Is comparable
		Complete data	Truncated data	
DFO-QUE-COD3Pn4RS-1964-2007	1	(–4.70, –0.22)	(–4.62, –0.72)	Yes
IMARPE-PANCHPERUNC-1963-2004	1	(–5.14, –0.80)	(–5.61, –1.25)	Yes
INIDEP-SBWHITARGS-1985-2007	1	(–10.67, –2.27)	(–14.46, –8.47)	Yes
NRIFS-OFLOUNECS-1986-2010	1	(–30.0, –14.2)	(–27.68, –2.50)	Yes
NWWG-HERRIsum-1984-2011	0	(–12.13, –3.53)	(–18.55, –2.62)	Yes
WGBFAS-HERR30-1972-2011	1	(–5.62, –0.16)	(–4.32, –0.09)	Yes
WGNSSK-WHITNS-VIId-IIIa-1989-2010	–2	(–10.23, –0.09)	(–7.60, 3.12)	No

Note: γ describes the best-fitted model, Complete Data describes the CI obtained for the complete population, Truncated Data describes the CI obtained after truncating the population at both ends, and Is Comparable indicates whether analysis repeated on truncated population agrees with the original one.

Table 7. Edge effect analysis applied to populations with more than 55 data points.

Assessment ID	γ	95% CI		Is comparable
		Complete data	Truncated data	
AFWG-HADNEAR-1947-2010	1	(–3.63, 0.52)	(–3.86, 0.11)	Yes
ICCAT-ATBTUNAEATL-1950-2010	0	(–4.62, 3.09)	(–3.88, 3.49)	Yes
NEFSC-HADGB-1930-2008	1	(–1.45, 2.59)	(–2.12, 2.14)	Yes
WGNSSK-CODNEAR-1943-2010	0	(–4.12, –0.36)	(–4.79, –0.88)	Yes

Note: γ describes the best fitted model, Complete Data describes the CI obtained for the complete population, Truncated Data describes the CI obtained after truncating the population at both ends, and Is Comparable indicates whether analysis repeated on truncated population agrees with the original one.

classified approximately the same number of populations, labelled with –1.

To adjust our results, we used our fitted models and tested whether we could reliably estimate the sign of η_1 with different confidence levels, as described in Table 5. For the case where the confidence level is 95%, we found seven populations labelled with –1, 82 populations labelled with 0, and one population labelled with +1. The entire classification list for the 95% confidence level is illustrated in Appendix A, Table A1.

Finally, we applied the edge effect analysis to populations classified with label –1 and to populations longer than 55 data points (Table A1). The former revealed an agreement in the classification of six of the seven populations so that we can assert that possible edge effects in the VPA are unlikely to be influencing the analysis for these model inputs (Table 6); however, in the latter we found a complete agreement with original results, indicating that the effect of VPA methods reduces with long time series populations (Table 7).

Discussion

This study develops, implements, and tests methods for identifying nonconstant variance (heteroscedasticity) in the spawner–recruit relationship. We found heteroscedastic models tend to fit the S-R model inputs better than constant variance models across the majority of stocks and found strong evidence for a negative coefficient of heteroscedasticity in seven cases (Table A1), including exploited cod, herring, and whiting stocks, in addition to olive flounder (*Paralichthys olivaceus*) and Peruvian anchoveta (*Engraulis ringens*). We advocate that the evidence for stochastic regulation in these cases deserves to be taken into account by managers. In contrast, only one stock was identified as having a positive coefficient of heteroscedasticity at the 95% confidence level.

We analysed the estimation error for η_1 (heteroscedasticity parameter) by exploring a class of heterogeneity models — frequentist and Bayesian paradigms — and associated model-fitting algorithms. Under the frequentist method, parameters are viewed as unknown but fixed quantities; consequently, the use of inferential procedures were evaluated under repeated sampling of the data. The frequentist method is generally easy to implement, but it encounters difficulties for small population size resulting in a large interval estimation and a loss of statistical significance. In contrast, Bayesian approaches can be appealing for problems of this sort, but difficulties arise in prior specification. Here, we used minimal prior information π_1 and π_2 to obtain the marginal posterior distribution for the parameters of interest; the estimation error for η_1 is obtained by estimating the Bayesian credible intervals using the posterior distribution.

To determine whether we can reliably estimate the sign of η_1, we tested whether the confidence interval falls in a region showing a consistent sign with the coefficient. We found that both frequentist and Bayesian methods led approximately to equivalent inference, but there are some circumstances under which one method outperforms the other, especially when the sample size is below 30 and when the Hessian matrix is not positive definite. The application of model selection reveals a consistent feature across all populations, as it selects a model having the best predictive ability among other models; in every case, heteroscedastic models fit the data better (i.e., lower AIC_c score), regardless of the sign of the coefficient of heteroscedasticity. This information is useful in a management context, where knowledge of the coefficient of heteroscedasticity is an important feature in assessing sustainable exploitation regimes (Minto et al. 2008; Burrow et al. 2013). This is illustrated in Appendix A (Table A1), which broadly labels each population from the set {–1, 0, +1}; the value –1 corresponds to stocks where there is good statistical evidence for a negative coefficient of heteroscedasticity (using, in this case, an approximate 95% confidence interval).

To reliably identify a negative coefficient of heteroscedasticity, managers or fisheries scientists using the frequentist methods should check that their chosen confidence interval lies in the negative region; those using the Bayesian framework can consider π_1 or π_2 as a noninformative benchmark prior and check whether their Bayesian credible interval lies in the negative region. We note that Bayesian approaches may be particularly useful where priors can be specified based on information about similar stocks in other locations. To protect this work against false positives or negatives, we recommend fisheries scientists to use both frequentist and Bayesian methods when assessing stocks for heteroscedasticity. If both meth-

ods agree, then there would be strong evidence that our conclusion is correct; otherwise, we should investigate the limitation of each method separately. To facilitate comparison, examples of the R code necessary for our analyses are supplied in the Supplementary material[1]. (The entire results can be reproduced by using the complete project built in R. Please see the file README.txt contained within the zip file of the Supplementary data for more details.)

Although both frequentist and Bayesian approaches were developed to identify the nonconstant variance exhibited in density-dependent models, heteroscedasticity could not be identified for the majority of the datasets no matter which method is used (out of 90 datasets, eight datasets are classified with label –1 and +1, under an approximate 95% confidence interval). The two principal reasons that drive this limited capacity to reliably identify heteroscedasticity are as follows: first, the data are typically rather poorly explained by the best-fitting S-R relationships because of the inherent noise in the S-R relationship; second, the time series are not long enough for reliable parameter estimation in most cases (Burrow et al. 2013). Furthermore, it is likely that, in some stocks, the magnitude of any heteroscedasticity is negligible. Nevertheless, this does not diminish the potential importance of heteroscedasticity and its identification, especially in the eight datasets for which we found good evidence of its presence.

We investigated whether there are natural clusterings of stocks with the same heteroscedastic classification; for example, one might hypothesize the same heteroscedastic signal of fish stocks of the same (or similar) species in different locations or alternatively in different stocks at the same location. Our preliminary analyses (using approximate 95% confidence levels) indicate no such convenient clusterings. However, further work is needed. For example, classification based on approximate 80% confidence levels reveals a consistent –1 classification for American plaice (*Hippoglossoides platessoides*), and such patterns may have relevance for sustainable management.

In this analysis, we made two assumptions. First, we discarded 10 S-R populations from the RAM legacy database because of missing data resulting in the analysis of 90 S-R populations of 32 species. We think that this sampling scheme had no bias implication because each population is treated individually and with no effect on the others. Second, we treated all VPA-type assessments as approximately equivalent, having first verified that the choice of an assessment had no statistical effect on the sign of coefficient of heteroscedasticity; this is validated by mapping the heteroscedasticity coefficient value against the VPA-type assessment where we found no impact of the VPA-type assessment on the classification method. Additionally, we assessed the possibility of edge effects in the VPA methods. Such effects may be caused by backward-convergence of VPA methods, increased variance of recruitment at the end of the time series, and shrinkage factors, and all these factors may introduce a bias in both SSB and recruitment estimates. This made no difference in the classification of 10 of the 11 populations tested, allowing us to confidently advocate the use of a heteroscedastic model with negative coefficient of heteroscedasticity as a valid management choice in these cases.

Our future work will seek to extend the analysis to a more holistic ecosystem-level analysis including external biotic and abiotic factors (i.e., an end-to-end perspective). Besides, we propose combining data from multiple stocks of similar species to better estimate the parameters of the S-R relationship. To account for heterogeneity, we propose considering a blocking factor and (or) within-block correlation in the log-likelihood function across different populations of similar species. If we do not get a blocking effect, meaning that two (or more) populations from the same species have the same variance, then pooling of multiple populations become statistically feasible. Alternatively, one can use data from multiple populations to obtain estimates of key parameters for individual populations throughout a Bayesian hierarchical framework (Gelman et al. 2004).

Acknowledgements

The authors gratefully acknowledge support from scientists managing the RAM Legacy Database. We especially acknowledge Daniel Ricard (Biology Centre AS CR v.v.i., Institute of Hydrobiology) and Olaf Jensen (Institute of Marine and Coastal Sciences, Rutgers University) for their guidance toward choosing the right S-R populations. We acknowledge Bradley Efron (Department of Statistics, Stanford University) for his guidance concerning the parametric bootstrap method. The authors also thank the anonymous reviewers for their valuable comments on this paper.

References

Akaike, H. 1973. Information theory and an extension of the maximum likelihood principle. *In* The Second International Symposium on Information Theory, Budapest. *Edited by* B.N. Petrov and F. Csaki. Akadémiai Kiado. pp. 267–281.

Anderson, D.E. 1978. An explanation of virtual population analysis. Vol. 19. National Marine Fisheries Service, Northeast Fisheries Center Woods Hole Laboratory, Woods Hole, Mass., USA.

Burrow, J.F., Horwood, J.W., and Pitchford, J.W. 2013. Variable variability: difficulties in estimation and consequences for fisheries management. Fish Fish. **14**(2): 205–212. doi:10.1111/j.1467-2979.2012.00463.x.

Daskalov, G. 1998. Using abundance indices and fishing effort data to tune catch-at-age analyses of sprat *Sprattus sprattus* L., whiting *Merlangius merlangus* L. and spiny dogfish *Squalus acanthias* L. in the Black Sea. *In* Dynamique des populations marines. pp. 215–228.

DiCiccio, T.J., and Efron, B. 1996. Bootstrap confidence intervals. Stat. Sci. **11**(3): 189–212. doi:10.1214/ss/1032280214.

Fournier, D.A., Skaug, H.J., Ancheta, J., Ianelli, J., Magnusson, A., Maunder, M.N., Nielsen, A., and Sibert, J. 2012. AD Model Builder: using automatic differentiation for statistical inference of highly parameterized complex nonlinear models. Optimization Methods and Software, **27**(2): 233–249. doi:10.1080/10556788.2011.597854.

Gelman, A., Carlin, J.B., Stern, H.S., and Rubin, D.B. 2004. Bayesian data analysis. *In* Press texts in statistical science. 2nd ed. Chapman & Hall/CRC.

Hilborn, R., and Walters, C.J. 1992. Quantitative fisheries stock assessment: Choice, dynamics, and uncertainty. Chapman and Hall, London, UK, and New York, USA.

Hsieh, C., Reiss, C.S., Hunter, J.R., Beddington, J.R., May, R.M., and Sugihara, G. 2006. Fishing elevates variability in the abundance of exploited species. Nature, **443**: 859–862. doi:10.1038/nature05232. PMID:17051218.

Hurvich, C.M., and Tsai, C.-L. 1989. Regression and time series model selection in small samples. Biometrika, **76**(2): 297–307. doi:10.1093/biomet/76.2.297.

Minto, C., Myers, R.A., and Blanchard, W. 2008. Survival variability and population density in fish populations. Nature, **452**(7185): 344–347. doi:10.1038/nature06605. PMID:18354480.

Peterman, R.M. 1981. Form of random variation in salmon smolt-to-adult relations and its influence on production estimates. Can. J. Fish. Aquat. Sci. **38**(9): 1113–1119. doi:10.1139/f81-151.

Pitchford, J.W., James, A., and Brindley, J. 2005. Quantifying the effects of individual and environmental variability in fish recruitment. Fish. Oceanogr. **14**(2): 156–160. doi:10.1111/j.1365-2419.2004.00299.x.

Plummer, M. 2003. JAGS: A program for analysis of Bayesian graphical models using Gibbs sampling. *In* Proceedings of the 3rd International Workshop on Distributed Statistical Computing, 25 January 2012, Vienna.

Ricard, D., Minto, C., Jensen, O.P., and Baum, J.K. 2012. Examining the knowledge base and status of commercially exploited marine species with the RAM Legacy Stock Assessment Database. Fish Fish. **13**(4): 380–398. doi:10.1111/j.1467-2979.2011.00435.x.

Shepherd, J.G., Cushing, D.H., and Beverton, R.J.H. 1990. Regulation in fish populations: myth or mirage? [and Discussion]. Philos. Trans. R. Soc. B Biol. Sci. **330**(1257): 151–164. doi:10.1098/rstb.1990.0189.

Singh, K. 1981. On the asymptotic accuracy of Efron's bootstrap. Ann. Stat. **9**(6): 1187–1195. doi:10.1214/aos/1176345636.

Steele, J. 1977. Fisheries mathematics: the proceedings of a conference. Academic Press.

Appendix A. Populations classification

Table A1 appears on the following pages.

Bibliography

Akaike, H. (1973). Information theory and an extension of the maximum likelihood principle. In Petrov, B. N. and Csaki, F., editors, *Second International Symposium on Information Theory*, pages 267–281, Budapest. Akadémiai Kiado.

Allenby, G., Rossi, P. E., and McCulloch, R. (2005). Hierarchical bayes model: A practitioners guide. *Journal of Bayesian Applications in Marketing*, pages 1–4.

Anderson, D. E. (1978). An explanation of virtual population analysis. *National Marine Fisheries Service.*

Andrieu, C., de Freitas, N., Doucet, A., and Jordan, M. I. (2003). An Introduction to MCMC for Machine Learning. *Machine Learning*, 50(1):5–43.

Barber, D. (2011). *Bayesian Reasoning and Machine Learning.* Cambridge University Press. In press.

Beddington, J. R. and May, R. M. (1977). Harvesting natural populations in a randomly fluctuating environment. *Science*, 197(4302):463–5.

Berger, J. O., Bernardo, J. M., and Sun, D. (2009). The formal definition of reference priors. *Annals of statistics*, 37:905.

Bernardo, J. and Smith, A. (2000). *Bayesian Theory.* John Wiley & Sons, New York. (ISBN: 0-471-92416-4).

Bernardo, J. M., Degroot, M. H., and Lindley, D. V. (1983). *Bayesian Statistics 2: Proceedings of the Second Valencia International Meeting.* Amsterdam: Elsevier Science Publishers B.V. ISBN 0-444-87746-0.

Beverton, R. J. H. and Holt, S. J. (1957). *On the dynamics of exploited fish populations*, volume 19. Chapman and Hall.

Bishop, C. M. (2007). *Pattern Recognition and Machine Learning*. Springer, 1st edition.

Bollerslev, T. (1986). Generalized Autoregressive Conditional Heteroskedasticity. *Journal of econometrics*, 31:307–327.

Bonnet, D., Harris, R., Yebra, L., Guilhaumon, F., Conway, D. V. P., , and Hirst, A. G. (2008). Temperature effects on *Calanus helgolandicus* (Copepoda: Calanoida) development time and egg production J. Plankton Res. *Journal of Plankton Research*, 31(1):31–44.

Box, G. and Jenkins, G. M. (1976). *Time series analysis: forecasting and control*. San Francisco: Holden Day, 1976.

Boyd, S. and Vandenberghe, L. (2004). *Convex optimization*. Cambridge University Press.

Brodziak, J. and Piner, K. (2010). Model averaging and probable status of North Pacific striped marlin, Tetrapturus audax. *Canadian Journal of Fisheries and Aquatic Sciences*, 67:793–805.

Buchel, C. and Friston, K. J. (1997). Characterizing functional integration. *Human Brain Function*, pages In R.S.J. Frackowiak, K.J Friston, C.D. Frith, R.J. Dolan, and J.C. Mazziotta, editors, Human Brain Function, pages 127–140.

Buckland, S. T., Newman, K. B., Thomas, L., and Koesters, N. B. (2004). State-space models for the dynamics of wild animal populations. *Ecological Modelling*, 171:157–175.

Burrow, J. F., Horwood, J. W., and Pitchford, J. W. (2012). Variable variability: difficulties in estimation and consequences for fisheries management. *Fish and Fisheries*, pages 205–212.

Byrd, R. H., Lu, P., Nocedal, J., and Zhu, C. (1995). A limited memory algorithm for bound constrained optimization. *SIAM J. Sci. Comput.*, 16(5):1190–1208.

Carvalho, C. M., Johannes, M. S., Lopes, H. F., and Polson, N. G. (2010). Particle Learning and Smoothing. *Statistical Science*, 25(1):88–106.

Castillo, E., Gutiérrez, J. M., and Hadi, A. S. (1997). *Expert Systems and Probabilistic Network Models*. Springer, New York, NY, USA, erste edition.

Chebichef, P. (1846). Dmonstration lmentaire d'une proposition gnrale de la thorie des probabilits. *Journal für die reine und angewandte Mathematik (Crelles Journal)*, 33:259–267.

Chen, Y., Chen, L., and Stergiou, K. I. (2003). Impacts of data quantity on fisheries stock assessment. *Aquat. Sci.*, 65:1–7.

Chen, Y. and Fournier, D. (1999). Impacts of a typical data on Bayesian inference and robust Bayesian approach in fisheries. *Canadian Journal of Fisheries and Aquatic Sciences*, 56(2):1525–1533.

Chiquet, J., Smith, A., Grasseau, G., Matias, C., and Ambroise, C. (2009). SIMoNe: Statistical Inference for MOdular NEtworks. *Bioinformatics*, 25:417–418.

Cox, R. T. (1946). Probability, Frequency and Reasonable Expectation. *American Journal of Physics*, 14(1):1.

Daskalov, G. (1998). Using abundance indices and fishing effort data to tune catch- at- age analyses of sprat Sprattus sprattus L., whiting Merlangius merlangus L. and spiny dogfish Squalus acanthias L. in the Black Sea. *Dynamique des populations marines*, pages 215–228.

de Finetti, B. (1931). Funzione caratteristica di un fenomeno aleatorio. *Academia Nazionale del Linceo.*

Deriso, R. (1980). Harvesting strategies and parameter estimation for an agestructured model. *Can. J. Fish. Aquat. Sci.*, 37:268–282.

DiCiccio, T. J. and Efron, B. (1996). Bootstrap confidence intervals. *Statistical Science*, 11(3):189–212.

Doucet, A., De Freitas, N., and Gordon, N., editors (2001). *Sequential Monte Carlo methods in practice*.

Doucet, A., Godsill, S., and Andrieu, C. (2000). On Sequential Monte Carlo Sampling Methods for Bayesian Filtering. *Statistics and computing*, 10(3):197–208.

Duane, S., Kennedy, A. D., Pendleton, B. J., and Roweth, D. (1987). Hybrid Monte Carlo. *Physics Letters B*, 195(2):216–222.

Edwards, M., Reid, P., and Planque, B. (2001). Long-term and regional variability of phytoplankton biomass in the northeast atlantic (19601995).

Efron, B. (1979). Bootstrap methods: Another look at the jackknife. *Annals of Statistics*, 7(3):1–26.

Efron, B., Hastie, T., Johnstone, I., and Tibshirani, R. (2004). Least angle regression. *The Annals of Statistics*, 32:407–499.

Efron, B. and Tibshirani, R. (1986). Bootstrap confidence intervals. *Statistical Science*, 1(1):54–77.

Engle, R. (1982). Autoregressive Conditional Heteroskedasticity With Estimates of the Variance of U.K. Inflation. In *Econometrica*, volume 50, pages 987–1008.

Fennel, W. (2010). A nutrient to fish model for the example of the baltic sea. *Journal of Marine Systems*, 81(1):184–195.

Fletcher, R. and Reeves, C. M. (1964). Function Minimization by Conjugate Gradients. *j-COMP-J*, 7(2):149–154.

Food and Agriculture Organisation (2016). The state of world fisheries and aquaculture 2016. contributing to food security and nutrition for all. rome. 200 pp. available online at `http://www.fao.org/3/a-i5555e.pdf`.

Forrest, R., McAllister, M., Dorn, M., Martell, S., and Stanley, R. D. (2010). Hierarchical bayesian estimation of recruitment parameters and reference points for pacific rockfishes (sebastes spp.) under alternative assumptions about the stock-recruit function. *Canadian Journal of Fisheries and Aquatic Sciences*, 67(10):1611–1634.

Fournier, D. A., Skaug, H. J., Ancheta, J., Ianelli, J., Magnusson, A., Maunder, M. N., Nielsen, A., and Sibert, J. (2012). AD Model Builder: using automatic differentiation for statistical inference of highly parameterized complex nonlinear models. *Optimization Methods and Software*, 27(2):233–249.

Friedman, J., Hastie, T., and Tibshirani, R. (2007). Sparse inverse covariance estimation with the graphical lasso. *Biostatistics*, 9:432–441.

Friel, N. and Pettitt, A., N. (2008). Marginal likelihood estimation via power posteriors. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 70(3):589–607.

Fromentin, J. M. and Planque, B. (1996). Calanus and environment in the eastern north atlantic .ii. influence of the north atlantic oscillation on c-finmarchicus and c-helgolandicus. *Mar. Ecol.-Prog. Ser.*, 134:111–118+.

Fulton, E. A. (2001). The effects of model structure and complexity on the behaviour and performance of marine ecosystem models. Technical report, University of Tasmania.

Fulton, E. A. (2010). Approaches to end-to-end ecosystem models. *Journal of Marine Systems*, 81:171–183.

Gelman, A. (2004). *Bayesian Data Analysis* . Chapman & Hall/CRC Press, 2nd ed. edition.

Gelman, A. and Meng, X. (1998). Simulating normalizing constants: from importance sampling to bridge sampling to path sampling. *Statistical Science)*, 13:163–185.

Gilks, W., Best, N., and Tan, K. (1995). Adaptive rejection metropolis sampling within gibbs sampling. *Applied statistics*, pages 455–472.

Gillespie, D. (1977). Exact stochastic simulation of coupled chemical reactions. *Journal of Physical Chemistry*, 81:2340–2361.

Godsill, S. J., Doucet, A., and West, M. (2004). Monte Carlo Smoothing for Nonlinear Time Series. *Journal of the American Statistical Association*, 99:156–168(13).

Gordon, N. J., Salmond, D. J., and Smith, A. F. M. (1993). Novel approach to nonlinear/non-Gaussian Bayesian state estimation. *Radar and Signal Processing IEE Proceedings F*, 140(2):107–113.

Hansen, B., Fotel, F., Jensen, N., and Madsen, S. (1996). Bacteria associated with a marine planktonic copepod in culture. ii. degradation of fecal pellets produced on a diatom, a nanoflagellate or a dinoflagellate diet. *Journal of Plankton Research*, 18(2):275–288.

Hastings, W. K. (1970). Monte Carlo sampling methods using Markov chains and their applications. *Biometrika*, 57(1):97–109.

Hawking, S. W. (1995). *A Brief History of Time*. Bantam.

Heisenberg, W. (1927). Über den anschaulichen Inhalt der quantentheoretischen Kinematik und Mechanik. *Zeitschrift für Physik A Hadrons and Nuclei*, 43:172–198.

Hering, D., Johnson, R., Kramm, S., Schmutz, S., Szoszkiewicz, K., and Verdonschot, P. (2006). Assessment of european streams with diatoms, macrophytes, macroinvertebrates and fish: a comparative metric-based analysis of organism response to stress. *Freshwater Biology*, 51(9):1757–1785.

Hilborn, R. and Walters, C. J. (1992). *Quantitative fisheries stock assessment choice, dynamics and uncertainty*, volume 2. Chapman and Hall.

Hill, S., Watters, G., Punt, A., McAllister, M., Qur, C., and Turner, J. (2007). Model uncertainty in the ecosystem approach to fisheries. *Fish and Fisheries*, 8(4):315–336.

Hjort, J. (1914). Fluctuations in the Great Fisheries of Northern Europe, Viewed in the Light of Biological Research. *Rapports et procès-verbaux des réunions*.

Hollowed, A. B., Barange, M., Ito, S.-I., Kim, S., Loeng, H., and Peck, M. e. (2011). Effects of climate change on fish and fisheries: forecasting impacts, assessing ecosystem responses, and evaluating management strategies. *ICES J. Mar. Sci.*, 68(6):983–1372.

Honaker, J. and King, G. (2010). What to do about missing values in time series cross-section data. *American Journal of Political Science*, 54(2):561–581.

Hsieh, C., Reiss, C. S., Hunter, J. R., Beddington, J. R., May, R. M., and Sugihara, G. (2006). Fishing elevates variability in the abundance of exploited species. *Nature*, 443:859–862+.

Hurrell, J. W. (1995). Decadal Trends in the North Atlantic Oscillation: Regional Temperatures and Precipitation. *Science*, 269(5224):676–679.

Hurvich, C. M. and Tsai, C. L. (1989). Regression and time series model selection in small samples. *Biometrika*, 76(21):297–307.

Irigoien, X., Harris, R. P., Head, R. N., and Harbour, D. (2000). North atlantic oscillation and spring bloom phytoplankton composition in the english channel. *Journal of Plankton Research*, 22(12):2367–2371.

Ito, S., Rose, K. A., Miller, A. J., Drinkwater, K., Brander, K., Overland, J. E., Sundby, S., Curchitser, E., Hurrell, J. W., and Yamanaka, Y. (2010). Ocean ecosystem responses to future global change scenarios: a way forward. in: Barange m., field j.g., harris r.h., hofmann e., perry r.i., werner f. (eds.). *Global Change and Marine Ecosystems*, pages 287–322.

Ives, A. and Carpenter, S. (2007). Stability and diversity of ecosystems. *Science*, 317:58–62.

Jeffreys, H. (1946). An Invariant Form for the Prior Probability in Estimation Problems. *Proceedings of the Royal Society of London. Series A, Mathematical and Physical Sciences*, 186(1007):453–461.

Johns, D. (2015). Monthly averaged data for fish eggs, fish larvae, total diatoms, total dinoflagellates, total large copepods and total small copepods as recorded by the continuous plankton recorder. *Sir Alister Hardy Foundation for Ocean Science. Plymouth.*

Kalman, R. E. (1960). A new approach to linear filtering and prediction problems. *Transactions of the ASME–Journal of Basic Engineering*, 82(Series D):35–45.

Keyl, F. and Wolff, M. (2008). Environmental variability and fisheries: what can models do? *Reviews in Fish Biology Fisheries*, 18:273–299.

Keynes, J. M. (1921). *A treatise on probability, by John Maynard Keynes.* Macmillan, London.

Koller, D. and Friedman, N. (2009). *Probabilistic Graphical Models: principles and techniques.* MIT press.

Kolmogorov, A. (1933). *Grundbegriffe der Wahrscheinlichkeitsrechnung.* Springer, Berlin.

LaMotte, L. (1983). Fixed-, random-, and mixed-effects models. *Encyclopedia of Statistical Sciences*, 3:137–141.

Lawrence, N., Girolami, M., Rattray, M., and Sanguinetti, G. (2010). *Learning and inference in computational systems biology*. MIT Press, Cambridge, MA.

Le Quéré, C., Moriarty, R., Andrew, R. M., Canadell, J. G., Sitch, S., Korsbakken, J. I., Friedlingstein, P., Peters, G. P., Andres, R. J., Boden, T. A., Houghton, R. A., House, J. I., Keeling, R. F., Tans, P., Arneth, A., Bakker, D. C. E., Barbero, L., Bopp, L., Chang, J., Chevallier, F., Chini, L. P., Ciais, P., Fader, M., Feely, R. A., Gkritzalis, T., Harris, I., Hauck, J., Ilyina, T., Jain, A. K., Kato, E., Kitidis, V., Klein Goldewijk, K., Koven, C., Landschützer, P., Lauvset, S. K., Lefèvre, N., Lenton, A., Lima, I. D., Metzl, N., Millero, F., Munro, D. R., Murata, A., Nabel, J. E. M. S., Nakaoka, S., Nojiri, Y., O'Brien, K., Olsen, A., Ono, T., Pérez, F. F., Pfeil, B., Pierrot, D., Poulter, B., Rehder, G., Rödenbeck, C., Saito, S., Schuster, U., Schwinger, J., Séférian, R., Steinhoff, T., Stocker, B. D., Sutton, A. J., Takahashi, T., Tilbrook, B., van der Laan-Luijkx, I. T., van der Werf, G. R., van Heuven, S., Vandemark, D., Viovy, N., Wiltshire, A., Zaehle, S., and Zeng, N. (2015). Global carbon budget 2015. *Earth System Science Data*, 7(2):349–396.

Lèbre, S. (2009). Inferring dynamic genetic networks with low order independencies. *Statistical Applications in Genetics and Molecular Biology*.

Leggett, W. C. and Deblois, E. (1994). Recruitment in marine fishes: Is it regulated by starvation and predation in the egg and larval stages? *Netherlands Journal of Sea Research*, 32(2):119 – 134.

Link, J. S., Fulton, E. A., and Gamble, R. J. (2010). The northeast us application of atlantis: A full system model exploring marine ecosystem dynamics in a living marine resource management context. *Progress in Oceanography*, 87:214–234.

Liu, R. (1988). Bootstrap procedures under some non i.i.d. models. *Annals of Statistics*, 16:1696–1708.

Ljung, G. M. and Box, G. E. P. (1978). On a measure of lack of fit in time series models. *Biometrika*, 65(2):297–303.

Lowe, R. M., Wu, W., Peterson, S. M., Brown-Peterson, J. N., Slack, T. W., and Schofield, J. P. (2012). Survival, growth and reproduction of non-native nile

tilapia ii: Fundamental niche projections and invasion potential in the northern gulf of mexico. *PLoS One*, 7(7).

Lunn, D., Jackson, C., Best, N., Thomas, A., and Spiegelhalter, D. (2013). *The BUGS Book: A Practical Introduction to Bayesian Analysis*. CRC Press.

Lyon, A. (2014). Why are normal distributions normal? *The British Journal for the Philosophy of Science*.

Mace, P. M. and Doonan, I. J. (1988). A generalized bioeconomic simulation model. *NZ fisheries assessment research*, 88/4.

MacKay, D. J. C. (2003). *Information Theory, Inference, and Learning Algorithms*. Copyright Cambridge University Press.

Malthus, T. (1798). *An essay on the principle of population*. Penguin Books, Middlesex.

McAllister, M. and Ianelli, J. (1997). Bayesian stock assessment using catch-age data and the sampling -importance resampling algorithm. *Canadian Journal of Fisheries and Aquatic Sciences*, 54:284–300.

McAllister, M., Pikitch, E., Punt, A., and Hilborn, R. (1994). A bayesian approach to stock assessment and harvest decisions using the sampling-importance resampling algorithm. *Canadian Journal of Fisheries and Aquatic Sciences*, 51:2673–2687.

McAllister, M. K. and Kirkwood, G. P. (1998). Applying multivariate conjugate priors in fishery-management system evaluation: how much quicker is it and does it bias the ranking of management options? *Canadian Journal of Fisheries and Aquatic Sciences*, 55(12):2642–2661.

Metropolis, N., Rosenbluth, A., Rosenbluth, M., Teller, A., and Teller, E. (1953). Equation of state calculations by fast computing machines. *Journal of Chemical Physics*, 21.

Michielsens, C., M.K., M., Kuikka, S., Mantyniemi, S., Romakkaniemi, A., Pakarinen, T., Karlsson, L., and Uusitalo, L. (2008). Combining multiple bayesian data analyses in a sequential framework for quantitative fisheries stock assessment. *Canadian Journal of Fisheries and Aquatic Sciences*, 65(5):962–974.

Michielsens, C. G. and McAllister, M. K. (2004). A Bayesian hierarchical analysis of stock-recruit data: quantifying structural and parameter uncertainties. *Canadian Journal of Fisheries and Aquatic Sciences*, 61(6):1032–1047.

Minto, C., Myers, R. A., and Blanchard, W. (2008). Survival variability and population density in fish populations. *Nature*, 452(7185):344–347.

Munch, S. B., Kottas, A., and Mangel, M. (2005). Bayesian nonparametric analysis of stock-recruitment relationships. *Canadian Journal of Fisheries and Aquatic Sciences*, 62(8):1808–1821.

Murphy, K. P. (2002). Dynamic bayesian networks: Representation, inference and learning. Technical report, UC Berkeley, Computer Science Division.

Myers, R. A. and Cadigan, N. G. (1993a). Density-dependent juvenile mortality in marine demersal fish. *Canadian Journal Of Fisheries And Aquatic Sciences*, 50.

Myers, R. A. and Cadigan, N. G. (1993b). Is juvenile natural mortality in marine demersal fish variable? *Canadian Journal of Fisheries and Aquatic Sciences*, 50(8):1591–1598.

Myers, R. A. and Mertz, G. (1998). Reducing uncertainty in the biological basis of fisheries management by meta-analysis of data from many populations: a synthesis. *Fisheries Research*, 37:51–60.

Nagarajan, R., Scutari, M., and Lèbre, S. (2013). Springer, New York.

Neal, R. (1993a). Bayesian learning via stochastic dynamics. In *Advances in Neural Information Processing Systems 5, [NIPS Conference]*, pages 475–482, San Francisco, CA, USA. Morgan Kaufmann Publishers Inc.

Neal, R. (1993b). Probabilistic inference using markov chain monte carlo methods.

Neal, R. (2003). Slice sampling. *Annals of statistics*, pages 705–741.

Neal, R. M. (1992). Bayesian Training of Backpropagation Networks by the Hybrid Monte Carlo Method. Technical report, Department of Computer Science, University of Toronto.

Neal, R. M. (1997). Monte Carlo Implementation of Gaussian Process Models for Bayesian Regression and Classification.

Needle, C. L. (2001). Recruitment models: diagnosis and prognosis. *Reviews in Fish Biology and Fisheries*, 11:95–111. 10.1023/A:1015208017674.

Nelder, J. A. and Mead, R. (1965). A simplex method for function minimization. *Computer Journal*, 7:308–313.

Newman, K. B. (2000). Hierarchic Modeling of Salmon Harvest and Migration. *Journal of Agricultural, Biological, and Environmental Statistics*, 5(4):pp. 430–455.

NOAA (2015). COPEPODITE: Interactive Time-series Explorer. Enter your Site's geographic location. `http://www.st.nmfs.noaa.gov/nauplius/media/html/subform_copepodite-c2.html`.

Opgen-Rhein, R. and Strimmer, K. (2007). From correlation to causation networks: a simple approximate learning algorithm and its application to high-dimensional plant gene expression data. *BMC Systems Biology*, 37(1).

Panikian, G., Cussens, J., and Pitchford, W., J. (2015). Identification and quantification of heteroscedasticity in stock-recruitment relationships. *Canadian Journal of Fisheries and Aquatic Sciences*, 72:1259–1271.

Parks, W. (1976). Cohort analysis, equilibrium yield per recruit analysis and predicted effects of minimum size limit regulation in the Atlantic bluefin tuna fisheries system. *ICCAT Coll.*, 5:313–331.

Peterman, R. (1981). Form of Random Variation in Salmon Smolt-to-Adult Relations and Its Influence on Production Estimates. *Canadian Journal of Fisheries and Aquatic Sciences*, 38(9):1113–1119.

Pitchford, J. W., James, A., and Brindley, J. (2005). Quantifying the effects of individual and environmental variability in fish recruitment. *Fisheries Oceanography*, 14(2):156–160.

Plummer, M. (2003). JAGS: A program for analysis of Bayesian graphical models using Gibbs sampling. In *Proceedings of the 3rd International Workshop on Distributed Statistical Computing*, Vienna.

Prado, R. and Lopes, H. F. (2010). Sequential parameter learning and filtering in structured autoregressive state-space models. *Statistics and Computing*, (2010):1–31.

Punt, A. E., Pribac, F., Walker, T., I., Taylor, B., and J.D., P. (2000). Stock assessment of school shark *galeorhinus galeus* based on a spatially explicit population dynamics model. *Mar. Freshw. Res.*, 51:205–220.

Rasmussen, C. E. and Williams, C. (2006). *Gaussian Processes for Machine Learning*. MIT Press.

Rattray, M. (2008). Probabilistic Inference and Learning: Master Course. *University of Manchester, U.K.*

Ricard, D., Minto, C., Jensen, O., and Baum, J. K. (2012). Evaluating the knowledge base and status of commercially exploited marine species with the ram legacy stock assessment database. *Fish and Fisheries*, 13:380–398.

Ricker, W. E. (1954). Stock and recruitment. *Journal of the Fisheries Research Board of Canada*, 11:559–623.

Robert, C. P. and Casella, G. (2009). *Introducing Monte Carlo Methods with R*. Springer Verlag, 1 edition.

Roberts, G. O. and Rosenthal, J. S. (1998). Optimal scaling of discrete approximations to langevin diffusions. *J. R. Statist. Soc. B*, 60:255–268.

Rose, K. A. (2012). End-to-end models for marine ecosystems: Are we on the precipice of a significant advance or just putting lipstick on a pig? *Scienta Marina*, 76(1):195–201.

Rose, K. A., Allen, J. I., Artioli, Y., Barange, M., Blackford, J., Carlotti, F., Cropp, R., Daewel, U., Edwards, K., Flynn, K., Hill, S. L., Hille, R., Lambers, R., Huse, G., Mackinson, S., Megrey, B., Moll, A., Rivkin, R., Salihoglu, B., Schrum, C., Shannon, L., Shin, Y.-J., Smith, S. L., Smith, C., Solidoro, C., St. John, M., and Zhou, M. (2010). End-to-end models for the analysis of marine ecosystems: challenges, issues, and next steps. *Marine and Coastal Fisheries: Dynamics, Management, and Ecosystem Science*, 2:115–130.

Rubin, D. B. (1988). Using the SIR algorithm to simulate posterior distributions. In Bernardo, M. H., Degroot, K. M., Lindley, D. V., and Smith, A. F. M., editors, *Bayesian Statistics 3*. Oxford University Press.

Rudolf, P., Serguei, S., and Jordi, B. (2014). On the structural stability of mutualistic systems. *Science*, 345:416–426.

Schafer, J., Opgen-Rhein, R., and Strimmer, K. (2006). Reverse engineering genetic networks using the genenet package. *R News*, 6:50–53.

Schnute, J. (1985). A general theory for analysis of catch and effort data. *Canadian Journal of Fisheries and Aquatic Sciences*, 42:414–429.

Schwarz, G. (1978). Estimating the dimension of a model. *Annals of Statistics*, 6:461–464.

Shao, J. (1988). On resampling methods for variance and bias estimation in linear models. *Annals of Statistics*, 16(3):986–1008.

Shepherd, J. G. and Cushing, D. H. (1990). Regulation in Fish Populations: Myth or Mirage? [and Discussion]. *Philosophical Transactions of the Royal Society B: Biological Sciences*, 330(1257):151–164.

Singh, K. (1981). On the Asymptotic Accuracy of Efron's Bootstrap. *Annals of Statistics*, 9(6):1187–1195.

Sinharay, S. and Stern, H. (2005). An empirical comparison of methods for computing bayes factors in generalized linear mixed models. *Journal of Computational and Graphical Statistics*, 14:415–435.

Spencer, P. D. (2008). Density-independent and density-dependent factors affecting temporal changes in spatial distributions of eastern Bering Sea flatfish. *Fisheries Oceanography*, 17(5):396–410.

Spiegelhalter, D., Best, N., Carlin, B., and van der Linde., A. (2002). Bayesian measures of model complexity and fit. *Journal of the Royal Statistical Society, Series B.*, 64:583–640.

Steele, J., of Mathematics, I., and Applications, I. (1977). *Fisheries mathematics: the proceedings of a conference*. Academic Press.

Stock, C., Alexander, M., Bond, N., Brander, K., Cheung, W., Curchitser, E. N., Delworth, T. L., Dunne, J. P., Griffies, S. M., Haltuch, M. A., Hare, J. A., Hollowed, A. B., Lehodey, P., Levin, S. A., Link, J. S., Rose, K. A., Rykaczewski, R. R., Sarmiento, J. L., Stouffer, R. J., Schwing, F. B., Vecchi, G. A., and Werner, F. E. (2011). On the use of ipcc-class models to assess the impact of climate on living marine resources. *Progress In Oceanography*, pages 88–27.

Stramer, O. and Tweedie, R. (1999). Langevin-type models ii: Self-targeting candidates for mcmc algorithms*. *Methodology And Computing In Applied Probability*, 1:307–328.

Sverdrup, H. U. (1953). On conditions for the vernal blooming of phytoplankton. *Journal du Conseil*, 18(3):287–295.

Tibshirani, R. (1994). Regression shrinkage and selection via the lasso. *Journal of the Royal Statistical Society, Series B*, 58:267–288.

Todhunter, I. (1865). *History of the mathematical theory of probability from the time of Pascal to that of Laplace*. The Nineteenth Century. title no. N.1.1.5855: General Collection. Macmillan and co.

Travers, M., Shin, Y., Jennings, S., and Cury, P. (2007). Towards end-to-end models for investigating the effects of climate and fishing in marine ecosystems. *Progress In Oceanography*, 75(4):751–770.

Travers, M., Shin, Y. J., Jennings, S., Machu, E., Huggett, J., Field, J., and Cury, P. (2009). Two-way coupling versus one-way forcing of plankton and fish models to predict ecosystem changes in the benguela. *Ecological Modelling*, 220(21):3089–3099.

Trifonova, N., Kenny, A., Maxwell, D., Duplisea, D., Fernandes, J., and Tucker, A. (2015). Spatio-temporal bayesian network models with latent variables for revealing trophic dynamics and functional networks in fisheries ecology. *Ecological Informatics*, 30:142–156.

Trifonova, N., Maxwell, D., Pinnegar, J., Kenny, A., and Tucker, A. (2017). Predicting ecosystem responses to changes in fisheries catch, temperature, and primary productivity with a dynamic bayesian network model. *ICES Journal of Marine Science: Journal du Conseil*.

Tsitsika, E. V., Maravelias, C. D., and Haralabous, J. (2007). Modeling and forecasting pelagic fish production using univariate and multivariate arima models. *Fisheries Science*, 73(5):979–988.

Turchin, P. (2001). Does Population Ecology Have General Laws? *Oikos*, 94:17–26.

Wakefield, J. (2013). *Bayesian and Frequentist Regression Methods*. Springer Series in Statistics.

Weber, R. (2007). *Time Series*. University of Cambridge. Lecture note.

Wilkinson, D. (2011). *Stochastic Modelling for Systems Biology, second edition*. Chapman & Hall/CRC.

Williams, C. and Rasmussen, C. (1996). Gaussian Processes for Regression. In *Advances in Neural Information Processing Systems 8*, pages 514–520. MIT press.

Wong, C. W. (2014). End-to-end models of marine ecosystems: exploring the consequences of climate change and fishing using a minimal framework. Technical report, University of York.

Wooster, W. W. and Bailey, K. M. (1989). Effects of ocean variability on recruitment and an evaluation of parameters used in stock assessment models. *Canadian Journal of Fisheries and Aquatic Sciences*, 108. Recruitment of marine fishes revisited. *In Beamish, R.J., and G.A. McFarlane (eds.).*

Wu, C. (1986). Jackknife, bootstrap and other resampling methods in regression analysis. Annals of Statistics*, 4:1261–1295.*