Prey Selection for Fatty Acid Signature Analysis Using the Akaike Information Criterion

by

Tyler Rideout

BSc Mathematics and Statistics, UNB, 2020

A THESIS SUBMITTED IN PARTIAL FULFILLMENT OF THE REQUIREMENTS FOR THE DEGREE OF

Master of Science in Statistics

In the Graduate Academic Unit of Mathematics and Statistics

Supervisor(s):	Matthew Stephenson, PhD, Department of Mathematics and Statistics Connie Stewart, PhD, Department of Mathematics and Statistics
Examining Board:	Maggie Brown, PhD, Department of Mathematics and Statistics Kimberley Davies, PhD, Department of Biological Sciences

This thesis is accepted by the
Dean of Graduate Studies

THE UNIVERSITY OF NEW BRUNSWICK

March, 2025

© Tyler Rideout, 2025

Abstract

The estimation of marine predator diet compositions through fatty acid signature analysis allows valuable insights into the trophic structures of marine ecosystems by comparing the fatty acid signatures of predators with a library of their potential prey. These prey libraries consist of individual fatty acid signatures grouped by species, and in order to obtain unique diet estimates the number of groups in the library must be less than the number of dietary fatty acids in the analysis. We propose a novel application of the Akaike information criterion to identify the correct subset of species from a wider set of potential prey. The estimation of true zeroes through the removal of species from the prey library may result in reduced variability and greater accuracy in the estimation of non-zero proportions. Outcomes from simulation studies as well as the analysis of real-life grey seal data are used to explore the performance of this method.

Acknowledgements

The foremost acknowledgement in this thesis must be of my supervisors, Dr. Connie Stewart and Dr. Matthew Stephenson, for whose support, guidance, and insights I am profoundly grateful. I have been exceedingly lucky to have had them supervising my progress, and truly value their time, commitment, and invaluable contributions.

I also extend my thanks to Dr. Andrea Burgess, whose meticulousness I admired as an undergraduate student and now deeply appreciate in a reviewer, and to Dr. Shelley Lang for providing the data that we relied upon in carrying out this work.

Table of Contents

Abstract ii

Acknowledgments iii

Table of Contents iv

List of Tables vii

List of Figures ix

1 Introduction 1

1.1 Diet Estimation for Marine Predators 1

1.2 Thesis Overview 3

2 Background 4

2.1 Compositional Data 4

2.1.1 Transformations 5

2.1.2 Measures of Distance 7

2.2 Fatty Acid Signature Analysis 9

2.2.1 QFASA 9

2.2.2 MUFASA 11

2.3 Variable Selection 14

3 Methods 16

3.1 Proposed Maximum Likelihood Diet Estimation Method 16

3.1.1 Maximum Likelihood Model for Diet Estimation 17
3.1.2 Computation of QFASA MLE Estimates 20
3.1.3 Aitchison's Distance Equivalence 20
3.2 Prey Selection 22
3.2.1 Backward Elimination 22
3.2.2 Forward Selection 23
3.3 Model Evaluation 25
3.3.1 Prey Library 25
3.3.2 Simulation Studies 28
3.3.3 Real-Life Example 31

4 Results **34**
4.1 Simulation Studies 34
4.1.1 Case 1: # Species $<$ # FAs 34
4.1.2 Case 2: # Species $\geq$ # FAs 41
4.2 Real-Life Example 46
4.2.1 Diet Estimates 46
4.2.1.1 Case 1: # Species $<$ # FAs 46
4.2.1.2 Case 2: # Species $\geq$ # FAs 48
4.2.2 Species Selection 51
4.2.2.1 Backward Elimination 51
4.2.2.2 Forward Selection 51
4.2.2.3 Candidate Models 53
4.2.2.4 Selection Differences Between Cases 54
4.3 Analysis of Prey Data 57

5 Discussion and Conclusions **64**
5.1 Discussion of Results 64

5.2 Future Work 67

Curriculum Vitae

List of Tables

3.1 Sample sizes and average lipid content as a percentage of total body weight for species included in the prey database. Species marked with an asterisk (*) appear only in the extended prey library. 26

3.2 List of the 67 distinct FAs measured from the prey samples. Those 28 FAs used for estimation form the dietary subset, grouped separately. 27

3.3 Dietary proportions used to generate pseudo-predators. * Ranges of proportions assigned to individual predators are indicated for Diet 4. 29

3.4 List of calibration coefficients (CCs) for the 28 dietary FAs. 32

4.1 Match rates (mean and SD as percentages) of the correct classification of true zero and true nonzero species by method for each diet. 39

4.2 Species-specific match rates of the correct classification of true zero species as a percentage of 100 iterations per diet. Dashes indicate rates that do not exist as the species is not a true zero in that diet. . 40

4.3 Match rates (mean and SD as percentages) of the correct classification of true zero and true nonzero species by FS for each diet using the extended prey library. 44

4.4 Species-specific match rates of the correct classification of true zero species as a percentage of 100 iterations per diet using the extended prey library. 45

4.5 Removal order of species added using backward elimination with the basal prey library (Case 1). 51

4.6 Selection order of species added using forward selection with the basal prey library (Case 1). Species marked with an asterisk are those that would have been included without a minimum model size imposed. . 52

4.7 Selection order of species added using forward selection with the extended prey library (Case 2). Species marked with an asterisk are those that would have been included without a minimum model size imposed. 53

4.8 Range of AIC values for the first three iterations of forward selection for seals sampled in 1997 and 2004 using basal (Case 1) and extended (Case 2) prey libraries. The number of candidate models whose AIC value is within 10% of the best is shown as a fraction of the total number of candidate models for that iteration. 54

4.9 Mean estimated dietary proportion assigned to select species by individuals of the same species using prey-on-prey analysis. 62

List of Figures

4.1 Diet estimates for 1500 individual pseudo-predators generated from Diet 1 (A) and Diet 2 (B). 35

4.2 Diet estimates for 1500 individual pseudo-predators generated from Diet 3 (A). Biases (estimated proportion - true proportion) on diet estimates for 1500 individual pseudo-predators generated from Diet 4 (B). 37

4.3 Chi-squared distances between true and estimated diet compositions for 1500 pseudo-predators generated from each diet. 38

4.4 Diet estimates for 1500 individual pseudo-predators generated from Diet 1 (A) and Diet 2 (B) following FS on the extended prey library compared against QFASA estimates using the exact subset of species included in the true diet. 42

4.5 Diet estimates for 1500 individual pseudo-predators generated from Diet 3 following FS on the extended prey library (A). Biases (estimated proportion - true proportion) on diet estimates for 1500 individual pseudo-predators generated from Diet 4 following FS on the extended prey library (B). Both are compared against QFASA estimates using the exact subset of species included in the true diet. . . . 43

4.6 Diet estimates for 39 adult grey seals sampled in 1997 using the basal prey library with a minimum model size of 10. Backward elimination (BE) and forward selection (FS) estimates are compared with QFASA estimates obtained using both Aitchison's distance (AIT) and Kullback-Leibler divergence (KL). 46

4.7 Diet estimates for 49 adult grey seals sampled in 2004 using the basal prey library with a minimum model size of 10. Backward elimination (BE) and forward selection (FS) estimates are compared with QFASA estimates obtained using both Aitchison's distance (AIT) and Kullback-Leibler divergence (KL). 47

4.8 Diet estimates for 39 adult grey seals sampled in 1997 using the extended prey library with a minimum model size of 10. Forward selection (FS) estimates are compared with QFASA estimates obtained using both Aitchison's distance (AIT) and Kullback-Leibler divergence (KL). 49

4.9 Diet estimates for 49 adult grey seals sampled in 2004 using the extended prey library with a minimum model size of 10. Forward selection (FS) estimates are compared with QFASA estimates obtained using both Aitchison's distance (AIT) and Kullback-Leibler divergence (KL). 50

4.10 Change in estimated dietary proportions obtained using forward selection with the basal (Case 1) and extended (Case 2) prey libraries. Values reflect the change from Case 1 to Case 2. 56

4.11 Clustering of species within the basal prey library as determined by Aitchison's (top panel) and KL (bottom panel) distances between mean FA signatures. 58

4.12 Clustering of species within the extended prey library as determined by Aitchison's and KL distances between mean FA signatures. 59
4.13 Mean prey-on-prey diet estimates for each species in the extended prey library using KL and Aitchison's distance. 61
4.14 Proportional representation of each species in the extended prey library in prey-on-prey diet estimates above or below the expected proportion of $1/33 = 0.03$. 63

Chapter 1

Introduction

1.1 Diet Estimation for Marine Predators

Knowledge of the diets of predators is crucial in understanding the dynamics of the environments they inhabit, an important goal in the field of ecology. While direct observation is a reasonable way to determine the feeding habits of some predators, it is less reasonable for others; in areas of marine ecology involving deep-diving predators, indirect methods must usually be relied upon to determine their diets. The analysis of feces and stomach contents has been used to reconstruct predator diets, but these methods are thought to result in biased estimates, accounting only for any digestion-resistant components of recently consumed prey (Iverson et al., 2004).

Methods relying upon the fatty acid signatures of predators and prey offer a quantitative alternative for diet estimation. Fatty acids (FAs) are the building blocks of lipids, and a subset of them, referred to as the dietary subset, are released from consumed lipids but not degraded during digestion (Cook, 2002); as such, when those particular FAs are liberated from the consumed prey they are absorbed into the predator's fat stores predictably and largely intact. In fatty acid signature analysis

(FASA), diet compositions are estimated using statistical models relating the FA signatures of predators with those of potential prey.

An important factor in these analyses is the choice of a prey library, the set of species whose FA signatures will be represented in the model and ultimately used to estimate what proportion the species comprises in the predator's diet. The number of potential species in a given environment may be vast, and the difficulty associated with determining the diets of some marine predators directly can pose a challenge in selecting a suitable prey library from a posteriori evidence. Nonetheless, some selection must take place, as the number of species in the prey library must not exceed the number of distinct FAs in the modelling set. Otherwise, the result is an undetermined system and estimates may not be unique (Phillips & Gregg, 2003).

There is little existing literature on statistical methods for prey selection in FASA. The issue has been previously explored in the context of estimating northern elephant seal diets with little prior knowledge of their prey in Goetsch et al. (2018), where the 51 potential prey species exceeded the 43 fatty acids in the dietary subset. To address the problem of non-unique estimates, the authors developed an ad hoc method for excluding species from the prey library. Original diet estimates were obtained using the full prey library, and 36 species found to be present in the estimated diet of at least one seal were identified as core prey. 36 new sets of diet estimates were obtained, each excluding one of the core prey species, and 10 of the remaining 15 species were ultimately excluded on the basis that they did not appear in proportions greater than 0.001 in any of the estimates. A review of the existing literature found no further publications of methods of prey selection.

This thesis proposes a maximum likelihood estimation (MLE) approach to FASA and with it, explores variable selection methods using the Akaike Information Criterion (AIC) in the context of determining which prey to include when obtaining diet estimates for marine predators. This work aims to establish an accessible sta-

tistical approach to reducing a library of potential prey to allow for estimation when there are more potential prey species than dietary FAs, and to determine whether more accurate diet estimates may be obtained when a prey selection algorithm is used.

1.2 Thesis Overview

Chapter 2 presents an overview of pertinent background information and is divided into three parts: the characteristics of compositional data and considerations necessary for its analysis, a review of existing methods for FASA, and a brief introduction to variable selection based on information criteria.

In Chapter 3, we propose an MLE model and show that when variance in the error is constant among the FAs, an equivalence can be drawn with Quantitative Fatty Acid Signature Analysis (QFASA), the most established and well-validated method of FASA (Iverson et al., 2004). We subsequently outline two variable selection methods using AIC: backward elimination (BE) and forward selection (FS). FS can be used to select a subset of prey in order to obtain valid diet estimates even when the number of potential prey species exceeds the number of dietary FAs.

Results of simulation studies using pseudo-predators generated from four known diets are presented in Chapter 4. These results explore the accuracy of diet estimates obtained by BE and FS, as well as the accuracy of species classification by both methods. A real-life example of prey selection carried out on two samples of grey seals follows. Throughout the chapter, results are compared against the benchmark of QFASA estimates. The chapter concludes with an analysis of the prey data.

Finally, Chapter 5 is comprised of a discussion section including considerations on our data, possible limitations of the proposed methods, ideas for future work in this area, and a conclusory summary of our key findings and recommendations.

Chapter 2

Background

2.1 Compositional Data

FA signatures can be thought of as vectors containing information about the proportions of individual FAs in a sample of fat tissue. In other words, the FA signature represents the FA composition of an individual's stored fat. The diet estimates rendered from these FA signatures represent the composition of a predator's diet, the proportion made up by individual prey species. The components of FASA are thus compositional in nature, belonging to a form of data whose analysis requires special consideration due to restrictions on the sample space. This section contains an overview of the characteristics of compositional data and those techniques applicable to its analysis in the setting of diet estimation, based largely on the work of Aitchison (1986).

Definition 2.1.1. *A composition* $\boldsymbol{x}$ *of* D *parts is a* $D \times 1$ *vector with non-negative components* $x_1, \ldots, x_D$ *subject to the constraint:*

$$x_1 + \ldots + x_D = 1.$$

This unit-sum constraint means that a D-part composition is completely spec-

ified by $d = D - 1$ of its components, since the D^{th} component is necessarily equal to the difference between the whole and the sum of the remaining d parts:

$$x_D = 1 - (x_1 + \ldots + x_d).$$

Definition 2.1.2. *The sample space for D-part compositions is the d-dimensional simplex defined by the set:*

$$S^d = \{(x_1, \ldots, x_D) \mid x_1 \geq 0, \ldots, x_D \geq 0,\ x_1 + \ldots + x_D = 1\}.$$

Definition 2.1.3. *The constraining operator $\mathscr{C}$ transforms a vector of non-negative components to a unit-sum vector, and is defined by:*

$$\mathscr{C}(\boldsymbol{x}) = \frac{1}{\Sigma_{i=1}^{D} x_i}\boldsymbol{x}.$$

Definition 2.1.4. *Let $\boldsymbol{x}$ be a D-part composition and $\boldsymbol{u}$ a vector of D non-negative elements. Then a perturbation is the operation defined by:*

$$\boldsymbol{u} \circ \boldsymbol{x} = \mathscr{C}(x_1 u_1, \ldots x_D u_D),$$

where $\boldsymbol{x}$ is perturbed by $\boldsymbol{u}$ and the resulting vector is also composition.

2.1.1 Transformations

Since compositions are made up of relative data and subject to a constant sum constraint, neither the components nor their covariances are free to vary independently, and the common assumption of multivariate normality is not applicable. Should standard techniques be applied in this setting, the risk for spurious correlation runs high. Transformations from the simplex to real space are often used in compositional data analysis.

Definition 2.1.5. *The centred log-ratio (clr) transformation is a symmetric transformation from the D-dimensional simplex $\mathcal{S}^D$ to $\mathcal{U}^D$, a D-dimensional hyperplane of real space, defined by:*

$$\boldsymbol{u}_c = clr(\boldsymbol{u}) = \left[\log \frac{u_1}{g(\boldsymbol{u})}, \ldots, \log \frac{u_D}{g(\boldsymbol{u})}\right]$$

where $g(\boldsymbol{u}) = (u_1 \cdots u_D)^{1/D}$ represents the empirical geometric mean of $\boldsymbol{u}$, and

$$\mathcal{U}^D = \{(u_1, u_2, \ldots, u_D) \mid u_1 + u_2 + \ldots + u_D = 0\}.$$

While the clr transformation preserves the original dimensionality of the data and eliminates the constraint of non-negativity, it replaces the unit-sum constraint with a zero-sum constraint and thus, a further transformation is required to address the remaining issue of collinearity.

Definition 2.1.6. *The isometric log-ratio (ilr) transformation maps from the D-dimensional simplex $\mathcal{S}^D$ to $\mathcal{R}^d$ and is defined by:*

$$\boldsymbol{u}^{ilr} = ilr(\boldsymbol{u}) = clr(\boldsymbol{u}) \cdot \boldsymbol{V}$$

where $\boldsymbol{V}$ is a $D \times d$ orthonormal basis of the clr-plane.

With the resultant sample space of $\mathcal{R}^d$, multivariate analysis techniques can be applied to ilr-transformed data.

The presence of zeroes in compositional data can be a barrier to the use of logarithms and ratios and thus, the above transformations. This may be addressed in various ways, one of which is the imputation technique defined as follows:

Definition 2.1.7. *The multiplicative replacement method is a technique for the im-*

putation of zeroes, calculated by:

$$r_j = \begin{cases} \delta_j & \text{if } x_j = 0 \\ \left(1 - \Sigma_{k|x_k=0}\delta_k\right) x_j & \text{if } x_j > 0\,, \end{cases}$$

where δ is a small positive value chosen to replace zeroes.

This technique is only appropriate for use with rounded zeroes, such as those in FA signatures, and not with essential zeroes indicating the true absence of a component such as those in diet estimates. Methodology for handling essential zeroes is discussed in Stewart and Field (2011), and in Tsagris and Stewart (2018).

2.1.2 Measures of Distance

A valid measure of distance between compositions is required as part of the optimization process to obtain diet estimates through FASA. Euclidean distance measures are not appropriate for compositions, which are not defined in Euclidean space as they do not satisfy its fundamental properties, but rather on the simplex (Aitchison, 1986). Several appropriate distance measures have been defined on the simplex.

Aitchison's distance is commonly used in the analysis of compositional data, and an equivalence between its use in QFASA and our proposed MLE approach to diet estimation is shown in Chapter 3. This measure is recommended for use in QFASA by Bromaghin et al. (2015) for its robustness to errors in calibration coefficients, a component of FASA models considered difficult to ascertain.

Definition 2.1.8. *Aitchison's distance between two D-part compositions* $\boldsymbol{x}_1$ *and* $\boldsymbol{x}_2$*, defined by Aitchison (1986), is calculated as:*

$$AIT(\boldsymbol{x}_1, \boldsymbol{x}_2) = \left(\sum_{j=1}^{D} \left(\log \left[\frac{x_{1j}}{\hat{g}(\boldsymbol{x}_1)} \right] - \log \left[\frac{x_{2j}}{\hat{g}(\boldsymbol{x}_2)} \right] \right)^2 \right)^{\frac{1}{2}},$$

where $\hat{g}(\boldsymbol{x}_i)$ *represents the empirical geometric mean of* $\boldsymbol{x}_i$*.*

The Kullback-Leibler (KL) distance has been widely used in FASA, beginning with its initial application to QFASA in Iverson et al. (2004), where it was found to perform favourably in giving weight to differences in rarer fatty acids.

Definition 2.1.9. *The Kullback-Leibler divergence between two compositions, defined by Kullback (1951), is calculated as:*

$$KL(\boldsymbol{x}_1, \boldsymbol{x}_2) = \sum_{j=1}^{D} (x_{1j} - x_{2j}) \log \left(\frac{x_{1j}}{x_{2j}} \right).$$

The chi-squared distance, proposed by Stewart (2016), may be an appropriate choice in the presence of zeroes as it is not reliant upon logarithms.

Definition 2.1.10. *The chi-squared distance between two D-part compositions is calculated as:*

$$CS(\boldsymbol{x}_1, \boldsymbol{x}_2) = \sqrt{2D} \left(\sum_{j=1}^{D} r_j \right)^{1/2},$$

where

$$r_j = \begin{cases} 0 & \text{if } x_{1j} = x_{2j} = 0 \\ \dfrac{\left(\frac{x_{1j}}{\Sigma_{k=1}^{D} x_{1k}} - \frac{x_{2j}}{\Sigma_{k=1}^{D} x_{2k}} \right)^2}{\frac{x_{1j}}{\Sigma_{k=1}^{D} x_{1k}} + \frac{x_{2j}}{\Sigma_{k=1}^{D} x_{2k}}} & \text{otherwise.} \end{cases}$$

2.2 Fatty Acid Signature Analysis

2.2.1 QFASA

Developed by Iverson et al. (2004), QFASA makes use of a statistical model informed by the properties of the storage and biosynthesis of FAs to produce quantitative estimates of the diet compositions of individual predators. This is done by minimizing the distance between the FA signature of the predator and a linear combination of representative FA signatures of the prey groups under consideration, where the representative FA signatures are taken to be the empirical sample means of each prey group. These prey groups are typically determined by species, but may include groups of similar species, or may conversely differentiate within-species clusters.

Taking I to be the number of distinct prey groups, let $\boldsymbol{\alpha} = \alpha_1, \ldots, \alpha_I$ denote the composition of a predator's diet, where the components represent the proportional contributions of each prey group to the diet. Let y_{jk} denote the proportion of the k^{th} FA, $k = 1, \ldots, D$, in the signature of the j^{th} predator, $j = 1, \ldots n$. Let x_{ilk} denote the proportion of the k^{th} FA in the signature of the l^{th} individual of the i^{th} species, and $\bar{\boldsymbol{x}}_i$ the mean FA signature for prey species i. Diet estimates $\boldsymbol{\alpha}_j$ for a single predator j are obtained by minimizing the distance, for some appropriate choice of distance measure, between the predator's FA signature $\boldsymbol{y}_j$ and $\hat{\boldsymbol{y}}_j$, where

$$\hat{\boldsymbol{y}}_j = \sum_{i=1}^{I} \alpha_i \, \bar{\boldsymbol{x}}_i. \tag{2.1}$$

Given the compositional nature of the outcome, this optimization is subject to the constraint that $\boldsymbol{\alpha}_j$ consists of non-negative components summing to 1. Diet estimates can be obtained using the QFASA R package (v1.2.1; Stewart et al., 2024).

The original choice to measure the distance between the compositions $\boldsymbol{y}_j$ and $\hat{\boldsymbol{y}}_j$ was the Kullback-Leibler divergence. This was considered by Iverson (2004) to

be a natural choice, as it is intended to measure the closeness between distributions, and FA signatures are akin to the distribution of FAs in a sample. This measure has since been used in a variety of validation studies. Aitchison's distance remains the most common measure of distance in compositional data analysis, and was found by Bromaghin (2015) to perform similarly and in some cases superiorly as a distance measure in QFASA. It should be noted that Aitchison's distance has several favourable properties, satisfying the criteria for compositional distance measures outlined by Aitchison (1992), but like the KL divergence, it is not an appropriate choice in the presence of essential zeroes.

Two biological corrections to the model were proposed by Iverson et al. (2004) to account for differences in fat content between prey species and in the metabolism of FAs. Since species with higher fat content contribute disproportionately to a predator's FA signature, a further correction is required to convert the proportional contribution of a species to the predator's FA signature to its proportion in the predator's diet. To account for the differences in fat content among prey species and the resultant impact on predator FA signatures, the diet estimates $\boldsymbol{\alpha}_i$ in Equation 2.1 are divided by the arithmetic means of the fat content for each species i and renormalized (Iverson et al., 2004).

The manner in which FAs are absorbed into a predator's fat stores is predictable, but not as simple as an exact mirroring of their prey's stores. To account for the metabolism of lipids, Iverson et al. (2004) developed the concept of calibration coefficients for use in FASA. These coefficients may be estimated through captive feeding studies, with predators fed a known diet for some period of time. For predators $j = 1, \ldots, n$, and individual prey $l = 1, \ldots, L$, $n \times L$ values r^k are calculated for each FA k as:

$$r^k_{lj} = \text{pred}_{jk}/\text{prey}_{lk}.$$

For each k, the highest and lowest 10% of values of r^k are trimmed before

taking the arithmetic mean of the remaining values to be the corresponding calibration coefficient c_k. The predator's FA proportions are divided by their corresponding calibration coefficients to account for the metabolism and biosynthesis of individual FAs, and the adjusted FA signature is renormalized. These adjusted FA signatures are used in the calculation of the diet estimates.

Recently, alternative methods have been proposed to estimate calibration coefficients simultaneously along with diet estimates (Bromaghin et al., 2017; McNichol, 2022). Methods for the computation of these coefficients are available in the QFASA R package.

2.2.2 MUFASA

Maximum Unified Fatty Acid Signature Analysis (MUFASA), developed by Steeves (2020), builds upon the QFASA model, differing in its use of maximum likelihood estimation and the treatment of prey as random effects, accounting for the fact that predators are not eating directly from the prey database. This method is intended to address QFASA's inability to account for the variability of FAs in prey signatures. MUFASA assumes that the ilr-transformed FA signatures follow a multivariate normal (MVN) distribution, allowing for variability in the prey; thus, the predator's FA signature can be considered a linear combination of the unobserved prey FAs perturbed by some error.

Let $\boldsymbol{Z}_{ji}$ denote the unobserved FA signature for prey species i consumed by predator j, and $\boldsymbol{\varepsilon}_j$ the random error associated with predator j. The predator's FA signature $\boldsymbol{y}_j$ is then given by:

$$\boldsymbol{y}_j = \left(\sum_{i=1}^{I} \alpha_i \, \boldsymbol{Z}_{ji} \right) \circ \boldsymbol{\varepsilon}_j. \tag{2.2}$$

It can be noted that the random effect $\boldsymbol{Z}_{ji}$ represents one unobserved prey

FA signature per species, and an alternative model discussed in Chapter 3 replaces it with the mean FA signature of the individual prey for each species. The choice to use a mean representative for each species may be supported by the idea that predators typically eat many individuals of the same species.

Let the superscript $^{\text{ilr}}$ denote data which have been ilr-transformed, $\boldsymbol{\Sigma}_\varepsilon$ the variance-covariance matrix of the random error, and let $\boldsymbol{\eta}_j = \sum_{i=1}^{I} \alpha_{ji}\, \boldsymbol{Z}_{ji}$. The likelihood function for a single predator j, easily extended to multiple predators as the product of the individual likelihoods, is given by Steeves (2020) as:

$$
\begin{aligned}
\mathcal{L}_j &= f(\mathbf{Y}_j^{ilr} | \mathbf{Z}_{j1}^{ilr}, \ldots, \mathbf{Z}_{jI}^{ilr}, \boldsymbol{\alpha}_j, \boldsymbol{\Sigma}, \boldsymbol{\Sigma}_\epsilon, \boldsymbol{\mu}_i, \mathbf{X}^{ilr}) \prod_{i=1}^{I} f(\mathbf{Z}_{ji}^{ilr}) \prod_{i=1}^{I} \prod_{m=1}^{n_i} f(\mathbf{X}_{im}^{ilr}) \\
&= \frac{1}{(2\pi)^{\frac{D-1}{2}} |\boldsymbol{\Sigma}_\epsilon|^{1/2}} \exp\left\{ -\frac{1}{2} (\mathbf{Y}_j^{ilr} - \boldsymbol{\eta}_j^{ilr})^T \boldsymbol{\Sigma}_\epsilon^{-1} (\mathbf{Y}_j^{ilr} - \boldsymbol{\eta}_j^{ilr}) \right\} \times \\
&\prod_{i=1}^{I} \frac{1}{(2\pi)^{\frac{D-1}{2}} |\boldsymbol{\Sigma}|^{1/2}} \exp\left\{ -\frac{1}{2} (\mathbf{Z}_{ji}^{ilr} - \boldsymbol{\mu}_i)^T \boldsymbol{\Sigma}^{-1} (\mathbf{Z}_{ji}^{ilr} - \boldsymbol{\mu}_i) \right\} \times \\
&\prod_{i=1}^{I} \prod_{m=1}^{n_i} \frac{1}{(2\pi)^{\frac{D-1}{2}} |\boldsymbol{\Sigma}|^{1/2}} \exp\left\{ -\frac{1}{2} (\mathbf{X}_{im}^{ilr} - \boldsymbol{\mu}_i)^T \boldsymbol{\Sigma}^{-1} (\mathbf{X}_{im}^{ilr} - \boldsymbol{\mu}_i) \right\}.
\end{aligned}
\tag{2.3}
$$

The inclusion of random effects in the model requires the integration of the joint likelihood with respect to the random effects $\boldsymbol{Z}_{ji}$ to obtain the marginal likelihood to be optimized. The negative log of the marginal likelihood is then optimized to find the most likely values for the diet composition $\boldsymbol{\alpha}$, again subject to the unit-sum and non-negativity constraints. To lessen the number of parameters to be optimized, this method estimates only four values for $\boldsymbol{\Sigma}_\varepsilon$, which is assumed to be a diagonal matrix. In practice, several steps must be carried out prior to the optimization. These steps, which are also used in the MLE method proposed in Chapter 3, are outlined by Steeves (2020) as follows:

1. Zeroes are modified using the multiplicative replacement method.

2. FA signatures are ilr-transformed.

3. Empirical sample means are calculated for each prey species, denoted by $\boldsymbol{\mu}_i$ in Equation 2.3.

4. A $(D-1) \times (D-1)$ estimated pooled covariance matrix is constructed from the ilr-transformed prey signatures, denoted by $\boldsymbol{\Sigma}$ in Equation 2.3.

5. QFASA estimates are used as starting values for the diet proportions, dropping the last proportion which can be calculated as 1 minus the sum of the estimated components.

6. The $(D-1) \times (D-1)$ diagonal covariance matrix of the ilr-transformed errors is reduced to four parameters representing the means of the quartiles of its diagonal, with starting values calculated as follows:

    1 For each predator, its diet composition is estimated using QFASA and used as the true diet in the generation of another 50 pseudo-predators without error.

    2 The estimates for $\boldsymbol{\varepsilon}_j$ are obtained by calculating $\text{ilr}((\Sigma_{i=1}^{I} \hat{\alpha}_{ij} \boldsymbol{x}_{ij}) \circ \boldsymbol{y}_j)$.

    3 The covariance matrix of the ilr-transformed errors is estimated.

    4 The diagonal values of the estimated covariance matrix are divided into quartiles and for each quartile, the mean of these values is calculated and used to represent the quartile, reducing the number of parameters to four.

Within the function p.MUFASA in the QFASA package in R, MUFASA goes on to obtain a marginal log-likelihood function using the package TMB (v1.9.15; Kristensen et al., 2024), which integrates out the prey random effects. This step is very computationally intensive, and the computational burden increases substantially as more prey species are included in the model.

2.3 Variable Selection

Statistical modelling aims to glean information about the relationship between a dependent variable and some number of independent variables (IVs). The choice of IVs considered in a model can be an important factor in its performance. Limitations on the number of IVs that can be included, often relative to the number of data points sampled, can necessitate the removal of some available candidate variables from consideration. Variable selection techniques aim to determine the correct subset of IVs to include in an analysis. While many such techniques exist, the majority are intended for use with data satisfying assumptions which are not met in the compositional setting.

In FASA, the variable selection problem corresponds to selecting the correct prey species to include in estimating the diets of predators. Existing research in this area is somewhat limited; an ad hoc prey selection method was developed and implemented by Goetsch et al. (2018) to obtain FASA diet estimates for northern elephant seals with little prior knowledge of their prey, evidencing both the lack of established methods and their potential utility.

With 51 potential prey species and 43 dietary FAs, Goetsch et al. (2018) obtained QFASA diet estimates using their full prey library and identified 36 species with estimated proportions > 0.001 in the diet of any individual seal as core prey. They then obtained QFASA diet estimates for the 36 models of size 50 excluding one core prey species at a time, and ultimately excluded 10 of the 15 non-core species based on the condition that they were not present in any of these sets of estimated diets, thus sufficiently reducing the size of the prey library to 41 species.

As a more generalized condition for variable selection, information criteria may be used to find the most suitable option among a set of plausible models. Using likelihood as a measure of goodness of model fit, the inclusion of additional IVs necessarily increases its measured fit (Heinze, 2017). Information criteria are used

to perform variable selection by adding a penalty term to the measure of fit that is related to the size of the model, or the number of IVs included. The increase in a model's likelihood achieved by the addition of an IV must thus be greater than the increased penalty it induces in order for it to be considered by the selection process to be a better model.

The Akaike Information Criterion (AIC) for a model, proposed by Akaike (1993), is calculated as:

$$AIC = -2 \log \mathcal{L} + 2 \cdot k, \tag{2.4}$$

where $\mathcal{L}$ is the maximum value of the likelihood function for the model and k is equal to the number of parameters in the model.

While we have assessed our method using the AIC in particular, other choices of information criteria exist, including the Bayesian Information Criterion (BIC) (Schwarz, 1978). These criteria function similarly, with the BIC differing only in the scale of the penalty term. It is calculated as:

$$BIC = -2 \log \mathcal{L} + \log(n) \cdot k, \tag{2.5}$$

where $\mathcal{L}$ is the maximum value of the likelihood function for the model, k is equal to the number of parameters in the model and n is the sample size.

In Chapter 3, we implement a variable selection method for prey using AIC to inform decisions in backward elimination and forward selection processes.

Chapter 3

Methods

3.1 Proposed Maximum Likelihood Diet Estimation Method

The specification of an appropriate likelihood model for diet estimation makes selection algorithms based on information criteria a viable option for determining a fitting subset of prey in FASA. Because these algorithms compare plausible models across various sizes, a single analysis requires the evaluation of numerous models. MUFASA models are thus rendered an infeasible choice in this setting due to the computational intensity induced by the treatment of prey as random effects. For computational ease, we propose an MLE model which omits the random effects and instead mirrors QFASA's use of mean representatives for each prey species in the equation for the assumed predator FA signatures. We justify the use of this model by showing an equivalence to QFASA under certain conditions. Henceforth, this model is referred to as the QFASA MLE method.

3.1.1 Maximum Likelihood Model for Diet Estimation

The components and indices referenced throughout this section follow the conventions established in the earlier discussion of existing methods where possible, and are summarized here.

Notation 3.1.1. *Indices:*

- $i = 1, \ldots, I$ *indexes prey species*
- $j = 1, \ldots, n$ *indexes individual predators*
- $k = 1, \ldots, D$ *indexes FAs*
- $l = 1, \ldots, L_i$ *indexes individual prey*

Notation 3.1.2. *Components:*

- $\boldsymbol{y}_j$ *denotes the D-dimensional FA signature of predator* j
- $\boldsymbol{x}_i$ *denotes the D-dimensional FA signature of a prey from species* i
- $\boldsymbol{\alpha}_j$ *denotes the I-dimensional diet composition of predator* j
- $\boldsymbol{\varepsilon}_j$ *denotes the D-dimensional random error associated with predator* j

Note that these terms are expressed on an untransformed compositional scale, and terms on the ilr-transformed scale are denoted by the addition of the ^{ilr} *superscript.*

As in MUFASA, the FA signatures of predators are assumed to be perturbed linear combinations of prey FA signatures. However, prey are represented by the arithmetic means of the species, as in QFASA, rather than random effects.

This assumed relationship is defined by the equation

$$\boldsymbol{y}_j = \left(\sum_{i=1}^{I} \alpha_{ji} \boldsymbol{\mu}_i \right) \circ \boldsymbol{\varepsilon}_j \tag{3.1}$$

where $\boldsymbol{\mu}_i$ is the mean FA signature for prey species i.

Further, the distributions of ilr-transformed components are assumed to be multivariate normal:

$$\begin{aligned} \boldsymbol{x}_i^{\text{ilr}} &\sim \text{N}^{D-1}(\boldsymbol{\mu}_i, \Sigma),\ i = 1, \ldots, I \\ \boldsymbol{\mathcal{E}}^{\text{ilr}} &\sim \text{N}^{D-1}(\mathbf{0}, \Sigma_\varepsilon), \end{aligned}$$

where $\boldsymbol{\mu}_i$ is the mean FA signature for species i, Σ is the common covariance matrix on the ilr-transformed scale and Σ_ε is a $D-1 \times D-1$ diagonal matrix.

Of the ilr-transformed predator FA signature $\boldsymbol{y}_j^{\text{ilr}}$, note that:

$$\begin{aligned} \boldsymbol{y}_j^{\text{ilr}} = \text{ilr}(\boldsymbol{y}_j) &= \text{ilr}\left[\left(\sum_{i=1}^{I} \alpha_{ji} \boldsymbol{\mu}_i \right) \circ \boldsymbol{\varepsilon}_j \right] \\ &= \text{ilr}\left(\sum_{i=1}^{I} \alpha_{ji} \boldsymbol{\mu}_i \right) + \text{ilr}(\boldsymbol{\varepsilon}_j) \end{aligned}$$

by the properties of the ilr transformation (Egozcue et al., 2003).

Let the ilr-transformed fitted predator FA signature be denoted by $\boldsymbol{\tau}_j^{\text{ilr}}$, where:

$$\boldsymbol{\tau}_j^{\text{ilr}} = \text{ilr}\left(\sum_{i=1}^{I} \alpha_{ji} \boldsymbol{\mu}_i \right).$$

Then, letting $\boldsymbol{x}^{\text{ilr}} = \left\{ \boldsymbol{x}_1^{\text{ilr}}, \ldots, \boldsymbol{x}_I^{\text{ilr}} \right\}$ and noting that the variance depends only on $\boldsymbol{\varepsilon}$ when conditioning on $\boldsymbol{X}^{\text{ilr}}$, we have:

$$\boldsymbol{Y}_j^{\text{ilr}} \mid \boldsymbol{X}^{\text{ilr}} \sim \text{N}^{D-1}\left(\boldsymbol{\tau}_j^{\text{ilr}}, \Sigma_\varepsilon \right).$$

We can assume that the errors $\boldsymbol{\varepsilon}_j$ associated with the predators are indepen-

dent of each other, the predators $\boldsymbol{y}_j$ are independent of each other, and the prey $\boldsymbol{x}_i$ are independent of each other. Given the assumed distributions and independences, the likelihood function for a single predator j is then:

$$\begin{aligned}
&\mathcal{L}(\boldsymbol{\alpha}_j, \Sigma_\varepsilon, \Sigma, \boldsymbol{\mu}_i \mid \boldsymbol{Y}_j^{\text{ilr}}, \boldsymbol{X}^{\text{ilr}}) = f(\boldsymbol{Y}_j^{\text{ilr}} \mid \boldsymbol{X}^{\text{ilr}}, \boldsymbol{\alpha}_j, \Sigma_\varepsilon, \Sigma, \boldsymbol{\mu}_i) \prod_{i=1}^{I} \prod_{m=1}^{n_i} f(\boldsymbol{X}_{im}^{\text{ilr}}) \\
&= \frac{1}{(2\pi)^{(D-1)/2} \mid \Sigma_\varepsilon \mid^{1/2}} \exp\left[-\tfrac{1}{2}(\boldsymbol{y}_j^{\text{ilr}} - \boldsymbol{\tau}_j^{\text{ilr}})^{\mathrm{T}} \Sigma_\varepsilon^{-1} (\boldsymbol{y}_j^{\text{ilr}} - \boldsymbol{\tau}_j^{\text{ilr}})\right] \times \\
&\prod_{i=1}^{I} \prod_{m=1}^{n_i} \frac{1}{(2\pi)^{(D-1)/2} \mid \Sigma \mid^{1/2}} \exp\left[-\tfrac{1}{2}(\boldsymbol{x}_{im}^{\text{ilr}} - \boldsymbol{\mu}_i)^{\mathrm{T}} \Sigma^{-1} (\boldsymbol{x}_{im}^{\text{ilr}} - \boldsymbol{\mu}_i)\right].
\end{aligned} \tag{3.2}$$

As in MUFASA, the $\boldsymbol{\mu}_i$ are estimated by $\bar{\boldsymbol{x}}_i$ and Σ is estimated by the pooled empirical variance-covariance matrices of the ilr-transformed prey (Steeves, 2020):

$$\begin{aligned}
&\mathcal{L}(\boldsymbol{\alpha}_j, \Sigma_\varepsilon, \Sigma, \boldsymbol{\mu}_i \mid \boldsymbol{Y}_j^{\text{ilr}}, \boldsymbol{X}^{\text{ilr}}) = f(\boldsymbol{Y}_j^{\text{ilr}} \mid \boldsymbol{X}^{\text{ilr}}, \boldsymbol{\alpha}_j, \Sigma_\varepsilon, \Sigma, \boldsymbol{\mu}_i) \prod_{i=1}^{I} \prod_{m=1}^{n_i} f(\boldsymbol{X}_{im}^{\text{ilr}}) \\
&= \frac{1}{(2\pi)^{(D-1)/2} \mid \Sigma_\varepsilon \mid^{1/2}} \exp\left[-\tfrac{1}{2}(\boldsymbol{y}_j^{\text{ilr}} - \hat{\boldsymbol{\tau}}_j^{\text{ilr}})^{\mathrm{T}} \Sigma_\varepsilon^{-1} (\boldsymbol{y}_j^{\text{ilr}} - \hat{\boldsymbol{\tau}}_j^{\text{ilr}})\right] \times \\
&\prod_{i=1}^{I} \prod_{m=1}^{n_i} \frac{1}{(2\pi)^{(D-1)/2} \mid \hat{\Sigma} \mid^{1/2}} \exp\left[-\tfrac{1}{2}(\boldsymbol{x}_{im}^{\text{ilr}} - \bar{\boldsymbol{x}}_i^{\text{ilr}})^{\mathrm{T}} \hat{\Sigma}^{-1} (\boldsymbol{x}_{im}^{\text{ilr}} - \bar{\boldsymbol{x}}_i^{\text{ilr}})\right],
\end{aligned} \tag{3.3}$$

where:

$$\hat{\boldsymbol{\tau}}_j^{\text{ilr}} = \text{ilr}\left(\sum_{i=1}^{I} \alpha_{ji} \bar{\boldsymbol{x}}_i\right).$$

Then the product in the 2^{nd} term is constant with respect to $\boldsymbol{\alpha}_j$ and Σ_ε, and the likelihood to be optimized over the diet estimates $\boldsymbol{\alpha}_j$ and covariance matrix Σ_ε for a single predator j is reduced to:

$$\begin{aligned}
&\mathcal{L}(\boldsymbol{\alpha}_j, \Sigma_\varepsilon \mid \boldsymbol{Y}_j^{\text{ilr}}, \boldsymbol{X}^{\text{ilr}}) \\
&\propto \frac{1}{(2\pi)^{(D-1)/2} \mid \Sigma_\varepsilon \mid^{1/2}} \exp\left[-\tfrac{1}{2}(\boldsymbol{y}_j^{\text{ilr}} - \hat{\boldsymbol{\tau}}_j^{\text{ilr}})^{\mathrm{T}} \Sigma_\varepsilon^{-1} (\boldsymbol{y}_j^{\text{ilr}} - \hat{\boldsymbol{\tau}}_j^{\text{ilr}})\right].
\end{aligned} \tag{3.4}$$

The likelihood function can be extended for multiple predators as the product of the individual likelihoods given in Equation 3.3. Then the function to be maximized with respect to the diet estimates $\boldsymbol{\alpha}_j$ and covariance matrix Σ_ε is:

$$\prod_{j=1}^{n} \mathcal{L}(\boldsymbol{\alpha}_j, \Sigma_\varepsilon \mid \boldsymbol{Y}_j^{\text{ilr}}, \boldsymbol{X}^{\text{ilr}}). \tag{3.5}$$

3.1.2 Computation of QFASA MLE Estimates

The process for obtaining diet estimates using the QFASA MLE method described above follows the process outlined in 2.2.2 through the calculation of starting values for Σ_ε. Computationally, it is more difficult to maximize Equation 3.4 than to equivalently minimize its negative log-likelihood (NLL). As such, the optimization used in the computation of the MLE estimates for $\boldsymbol{\alpha}_j$ is the minimization of the NLL using the R package Rsolnp. Recalling that the $\boldsymbol{\alpha}_j$ are compositions, the optimization is subject to the constraint

$$\sum_{i=1}^{I-1} \alpha_{ji} \leq 1 \, .$$

The constraint is so written as the values for α_{jI} are excluded from the optimization and calculated later as $1 - \Sigma_{i=1}^{I-1} \alpha_{ji}$. Further, the individual diet parameter estimates are bounded between 0 and 1.

3.1.3 Aitchison's Distance Equivalence

To justify the use of this method, we note its equivalence to QFASA (Equation 2.1) when Aitchison's distance measure is used and constant variance is assumed among the error associated with individual FAs. The MLE estimates for $\boldsymbol{\alpha}_j$ are obtained by simultaneously solving the equations

$$\frac{\partial \log \mathcal{L}(\boldsymbol{\alpha}_j, \Sigma_\varepsilon)}{\partial \boldsymbol{\alpha}_j} = \mathbf{0}\,, j = 1, \ldots, n\,, \tag{3.6}$$

while the QFASA estimates for $\boldsymbol{\alpha}_j$ using Aitchison's distance measure are obtained by solving the equations

$$\frac{\partial \mathrm{AIT}(\boldsymbol{y}_j, \hat{\boldsymbol{\tau}}_j)}{\partial \boldsymbol{\alpha}_j} = \mathbf{0}\,, j = 1, \ldots, n\,. \tag{3.7}$$

If the errors associated with the fatty acids, $k = 1, \ldots, D$, are assumed to have constant variance, such that $\sigma_1^2 = \ldots = \sigma_D^2 = \sigma^2$, then the optimizations (3.5) and (3.6) are equivalent.

Proof. The MLE estimates rely on the partial derivative of the log of the likelihood function with respect to $\boldsymbol{\alpha}_j$. Assume that $\sigma_1^2 = \sigma_2^2 = \ldots = \sigma_D^2 = \sigma^2$.

$$\begin{aligned}
\log \mathcal{L}(\boldsymbol{\alpha}_j, \Sigma_\varepsilon) &= \log\left(\frac{1}{(2\pi)^{(D-1)/2} \mid \Sigma_\varepsilon \mid^{1/2}} \exp\left[-\tfrac{1}{2}(\boldsymbol{y}_j^{\mathrm{ilr}} - \hat{\boldsymbol{\tau}}_j^{\mathrm{ilr}})^{\mathrm{T}} \Sigma_\varepsilon^{-1} (\boldsymbol{y}_j^{\mathrm{ilr}} - \hat{\boldsymbol{\tau}}_j^{\mathrm{ilr}})\right]\right) \\
&= -(\tfrac{D-1}{2}) \log(2\pi) - \tfrac{1}{2} \log(\mid \Sigma_\varepsilon \mid) - \tfrac{1}{2}[(\boldsymbol{y}_j^{\mathrm{ilr}} - \hat{\boldsymbol{\tau}}_j^{\mathrm{ilr}})^{\mathrm{T}} \Sigma_\varepsilon^{-1} (\boldsymbol{y}_j^{\mathrm{ilr}} - \hat{\boldsymbol{\tau}}_j^{\mathrm{ilr}})] \\
\frac{\partial \log \mathcal{L}(\boldsymbol{\alpha}_j, \Sigma_\varepsilon)}{\partial \boldsymbol{\alpha}_j} &= -\frac{1}{2}\left(\frac{\partial}{\partial \boldsymbol{\alpha}_j}\left[\left(\boldsymbol{y}_j^{\mathrm{ilr}} - \hat{\boldsymbol{\tau}}_j^{\mathrm{ilr}}\right)^{\mathrm{T}} \Sigma_\varepsilon^{-1} (\boldsymbol{y}_j^{\mathrm{ilr}} - \hat{\boldsymbol{\tau}}_j^{\mathrm{ilr}})\right]\right) \\
&= -\frac{1}{2}\left(\frac{\partial}{\partial \boldsymbol{\alpha}_j}\left[\sum_{k=1}^{D}\left(\left(y_{jk}^{\mathrm{ilr}} - \hat{\tau}_{jk}^{\mathrm{ilr}}\right)^2 \frac{1}{\sigma_k^2}\right)\right]\right) \\
&= -\frac{1}{2\sigma^2}\left(\frac{\partial}{\partial \boldsymbol{\alpha}_j}\left[\sum_{k=1}^{D}\left(\left(y_{jk}^{\mathrm{ilr}} - \hat{\tau}_{jk}^{\mathrm{ilr}}\right)^2\right)\right]\right) \text{ (since } \sigma_1^2 = \ldots \sigma_D^2 = \sigma^2) \\
&= -\frac{1}{2\sigma^2}\left(\frac{\partial}{\partial \boldsymbol{\alpha}_j}\left[\mathrm{AIT}(\boldsymbol{y}_j, \hat{\boldsymbol{\tau}}_j)\right]^2\right)
\end{aligned}$$

Recall from (3.5) that the MLE estimates for $\boldsymbol{\alpha}_j$ are obtained by setting this derivative to 0, thus minimizing Aitchison's distance as in QFASA.

$$\begin{aligned}
0 &= -\frac{1}{2\sigma^2}\left(\frac{\partial}{\partial \boldsymbol{\alpha}_j}\left[\mathrm{AIT}(\boldsymbol{y}_j, \hat{\boldsymbol{\tau}}_j)\right]^2\right) \\
\rightarrow 0 &= \frac{\partial}{\partial \boldsymbol{\alpha}_j}\left[\mathrm{AIT}(\boldsymbol{y}_j, \hat{\boldsymbol{\tau}}_j)^2\right]\,.
\end{aligned}$$

□

3.2 Prey Selection

Given the likelihood function in Equation 3.4, the information criterion given in Equation 2.4 may be used to select the best of a set of plausible models. This means selecting the best fitting subset of prey species as determined by the value of an information criterion, a measure of fit penalized proportionately to the number of species included. Because the negative term in the AIC measures the apparent model fit while the positive term is the penalty associated with the model size, better models are associated with lower AIC values. While this criterion can be used to select models in various ways, we have focused on the backward elimination and forward selection algorithms outlined below.

For both, consider a full potential prey library comprised of I distinct species. Let $\mathcal{M}_i^m$ denote model i of size m. In keeping with previous notation, we can think of i as indexing the individual species being added to or dropped from the previous model. Let $\text{AIC}_{\mathcal{M}_i^m}$ denote the AIC value associated with $\mathcal{M}_i^m$, and let $\text{AIC}^*_{\mathcal{M}^m}$ denote the lowest AIC value associated with a model of size m. Let i^* denote the species associated with the lowest AIC for a given model size. Let t index the iterations of the algorithm, beginning at $t = 1$ and increasing by 1 with each repetition of steps.

3.2.1 Backward Elimination

Backward elimination begins with the full model including all species in the prey library and removes species one at a time until a further reduction would cease to improve the AIC value. The algorithm for backward elimination proceeds as follows:

1. Obtain MLE estimates for the full model $\mathcal{M}^I$ and calculate $\text{AIC}_{\mathcal{M}^I}$.

2. Obtain MLE estimates for the $\mathcal{M}_i^{I-t}$, all models of size $I-t$ which are possible by excluding one species from the prey library, and calculate the $\text{AIC}_{\mathcal{M}_i^{I-t}}$. Determine $\text{AIC}^*_{\mathcal{M}^{I-t}} = \min_i(\text{AIC}_{\mathcal{M}_i^{I-t}})$.

3. If $\text{AIC}^*_{\mathcal{M}^{I-t}} = \text{AIC}_{\mathcal{M}_{i^*}^{I-t}} < \text{AIC}^*_{\mathcal{M}^{I-t+1}}$, remove species i^* from the prey library. Increase t to $t+1$.

4. Repeat steps 2 and 3 until $\text{AIC}^*_{\mathcal{M}^m} \geq \text{AIC}^*_{\mathcal{M}^{m+1}} = \text{AIC}_{\mathcal{M}_{i^*}^{m+1}}$.

5. Let $\mathcal{M}_{i^*}^{m+1}$ be the chosen model.

In words, on each iteration we evaluate a candidate model for each species remaining in the prey library by removing that one species from the current model. If any of these candidates result in a lower AIC value than the current model, we choose the best among them to replace it and repeat the process for as long as the AIC continues to improve.

As a cautionary measure to avoid excluding any species consumed by at least one individual predator, we include an option to restrict the list of species considered for removal from the model at each step. Let c be a specified threshold for a species to be considered for removal. At the beginning of each iteration, prior to carrying out step 2, QFASA diet estimates are obtained for each predator with the function p.QFASA from the QFASA package using the current prey library of $I-t+1$ species. If the QFASA estimate $\alpha_{ji} \geq c$ for any predator j and species i, species i is not considered for removal.

3.2.2 Forward Selection

In unconstrained data analysis, forward selection algorithms may begin with an empty model, adding one explanatory variable at a time until a further addition

would cease to improve the AIC value. Given the compositional nature of the outcome variable $\boldsymbol{\alpha}$, beginning with just one species i would necessarily result in an estimate of $\alpha_i = 1$ and the likelihood cannot be computed. We propose two options for addressing this:

1. A subset of at least two starting species may be specified based on prior knowledge.

2. As determined by AIC values, the best of all $\binom{I}{2}$ possible models of size 2 may be chosen as the starting subset.

Assume that $s \geq 2$ starting species are selected by either option and let the starting prey library be comprised of only those species. Then the forward selection algorithm proceeds as follows:

1. Obtain MLE estimates for the chosen $\mathcal{M}^s$ and calculate $\text{AIC}_{\mathcal{M}^s}$.

2. Obtain MLE estimates for the $\mathcal{M}_i^{s+t}$, all models of size $s+t$ which are possible by adding one species to the prey library, and calculate the $\text{AIC}_{\mathcal{M}_i^{s+t}}$. Determine $\text{AIC}^*_{\mathcal{M}^{s+t}} = \min_i(\text{AIC}_{\mathcal{M}_i^{s+t}})$.

3. If $\text{AIC}^*_{\mathcal{M}^{s+t}} = \text{AIC}_{\mathcal{M}_{i^*}^{s+t}} < \text{AIC}^*_{\mathcal{M}^{s+t-1}}$, add species i to the prey library. Increase t to $t+1$.

4. Repeat steps 2 and 3 until $\text{AIC}^*_{\mathcal{M}^m} \geq \text{AIC}^*_{\mathcal{M}^{m-1}} = \text{AIC}_{\mathcal{M}_{i^*}^{m-1}}$.

5. Let $\mathcal{M}_{i^*}^{m-1}$ be the chosen model.

In words, on each iteration we evaluate a candidate model for each potential species not yet included by adding that one species to the current model. If any of these candidates result in a lower AIC value than the current model, we choose the best among them to replace it and repeat the process for as long as the AIC continues to improve.

The application of these prey selection algorithms is divided into two cases which must be considered separately, and are referenced throughout the remainder of this thesis. Let K and I be the number of distinct FAs and prey species in the analysis, respectively. Then the cases to be considered are:

Case 1: $I < K$ (fewer species than FAs)

Case 2: $I \geq K$ (more species than FAs).

It is impossible to obtain unique, and thus valid, QFASA estimates in Case 2. When the true diet of predators is unknown and the number of potential prey species exceeds the number of dietary FAs, the implementation of a sound prey selection algorithm may be especially valuable in obtaining estimates where it would not otherwise be possible. While either algorithm may be applied in Case 1, only forward selection is appropriate for use in Case 2. Both cases are explored in Chapter 4.

3.3 Model Evaluation

3.3.1 Prey Library

To evaluate the prey selection algorithm in the first case (fewer prey species than FAs), a basal prey library comprised of FA signatures for 1689 individuals sampled among 21 distinct species is chosen. This group of species has previously been used in diet estimation for grey seals based on prior knowledge of their feeding habits as outlined in Stewart (2022). Sample collection and processing procedures are described in Budge et al. (2002).

In the second case (fewer FAs than prey species), an extended prey library is formed through the addition of FA signatures for 453 individuals among a further 12 species to the basal library for a total of 33 distinct species (Table 3.1).

Table 3.1: Sample sizes and average lipid content as a percentage of total body weight for species included in the prey database. Species marked with an asterisk (*) appear only in the extended prey library.

Species	**Scientific Name**	**n**	**Lipid (%)**
American lobster*	*Homarus americanus*	21	1.5
American plaice	*Hippoglossoides platessoides*	134	2.4
Atlantic argentine*	*Argentina silus*	25	5.1
Atlantic butterfish	*Peprilus triacanthus*	75	10.8
Atlantic cod	*Gadus morhua*	109	2.5
Atlantic herring	*Clupea harengus*	229	6.4
Atlantic mackerel	*Scomber scombrus*	32	5.1
Capelin	*Mallotus villosus*	162	5.2
Gaspereau*	*Alosa pseudoharengus*	70	9.6
Haddock*	*Melanogrammus aeglefinus*	115	1.7
Longhorn sculpin	*Myoxocephalus octodecemspinosus*	45	2.1
Marlin-spike grenadier*	*Nezumia bairdii*	5	3.2
Monkfish*	*Lophius americanus*	9	0.8
Northern sandlance	*Ammodytes dubius*	148	5.3
Northern shortfin squid	*Illex illecebrosus*	35	3.0
Northern shrimp*	*Pandalus borealis*	110	2.2
Ocean pout*	*Zoarces americanus*	31	2.3
Pollock	*Pollachius virens*	53	2.4
Red hake*	*Urophycis chuss*	18	1.5
Redfish	*Sebastes sp.*	54	7.1
Sea raven	*Hemitripterus americanus*	71	2.0
Silver hake	*Merluccius bilinearis*	58	1.6
Smooth skate	*Malacoraja senta*	33	2.5
Snake blenny	*Lumpenus lumpretaeformis*	18	2.4
Snow crab*	*Chionoecetes opilio*	25	0.9
Thorny skate	*Amblyraja radiata*	83	2.6
Toad crab*	*Hyas sp.*	12	0.6
White hake	*Urophycis tenuis*	80	1.3
Windowpane flounder*	*Scophthalmus aquosus*	12	1.8
Winter flounder	*Pseudopleuronectes americanus*	50	2.0
Winter skate	*Leucoraja ocellata*	40	1.5
Witch flounder	*Glyptocephalus cynoglossus*	24	1.9
Yellowtail flounder	*Limanda ferruginea*	156	2.2

Six of the species added to the extended prey library (American lobster, gaspereau, haddock, northern shrimp, ocean pout, and red hake) have been included previously in the estimation of grey seal diets (Beck et al., 2007), as permitted by the use of a larger subset of FAs for analysis. These species, thought only to appear in trace amounts in the diets of grey seals, were subsequently selected for removal when the recommended dietary subset of FAs was reduced. The final six species in the prey database were selected on the basis of having been sampled within the same general location (Northwest Atlantic Fisheries Organization Subarea 4) and season (May - June) as the others.

The total number of FAs measured in the prey signatures is 67, with 28 of them forming the dietary subset used for diet estimation (Table 3.2). Thus, the basal prey library of 21 species has fewer species than FAs (Case 1) and the extended prey library of 33 species has more species than FAs (Case 2).

Table 3.2: List of the 67 distinct FAs measured from the prey samples. Those 28 FAs used for estimation form the dietary subset, grouped separately.

Dietary Subset		**Additional Measured FAs**		
16:2n-6	20:1n-9	12:0	16:1n-11	18:1n-7
16:2n-4	20:1n-7	13:0	16:1n-9	18:1n-5
16:3n-6	20:2n-6	Iso-14	16:1n-7	18:2d-11
16:3n-4	20:3n-6	14:0	7Me16:0	18:2n-7
16:4n-1	20:4n-6	14:1n-9	16:1n-5	20:0
18:2n-6	20:3n-3	14:1n-7	Iso-17	20:2n-9
18:2n-4	20:4n-3	14:1n-5	17:0	22:1n-11
18:3n-6	20:5n-3	Iso-15	17:1	22:1n-9
18:3n-4	21:5n-3	Anti-15	16:3n-1	22:1n-7
18:3n-3	22:4n-6	15:0	16:4n-3	22:2n-6
18:3n-1	22:5n-6	15:1n-8	18:0	24:1n-9
18:4n-3	22:4n-3	15:1n-6	18:1n-13	
18:4n-1	22:5n-3	Iso-16	18:1n-11	
20:1n-11	22:6n-3	16:0	18:1n-9	

3.3.2 Simulation Studies

First outlined in Iverson et al. (2004), FASA is generally assessed through simulation studies using pseudo-predators whose FA signatures are generated based upon a specified diet. Four "true diet" compositions were chosen for simulations (Table 3.3).

The species assigned nonzero dietary proportions ("true nonzero") were selected as realistic prey species for grey seals (Beck et al., 2007). For each of the first three diets, compositions were constant and individual pseudo-predators within a sample were generated using identical proportions. Diet 1 was comprised of five prey species in unequal amounts, Diet 2 was comprised of five prey species in equal amounts, and Diet 3 was comprised of ten prey species in unequal amounts. Diet 4 was comprised of five prey species with different proportions assigned to each individual pseudo-predator within a sample.

For each of the true diets, 100 samples of 15 pseudo-predators were generated non-parametrically (i.e., with no parametric distributional assumptions being made) using the pseudo.pred function from the QFASA package in R. Using this method, individuals from the prey database were sampled with replacement within species and the mean species signature contributed in the appropriate proportion for each species to the pseudo-predator's FA signature. Each FA signature was then perturbed by a random error term to increase variability within the samples (Steeves, 2020). No adjustments for lipid content or calibration coefficients were made, and so resulting pseudo-predators may be less variable than in practice.

Beginning with the basal prey library, for each sample, diet compositions were estimated for each of the 15 pseudo-predators using QFASA. For each diet, BE and FS were performed on each of the 100 samples and the MLE diet estimates using the selected prey subset were obtained for each. BE used a threshold of > 0.01 in the QFASA [AIT] diet estimates of all predators for a species to be considered for

Table 3.3: Dietary proportions used to generate pseudo-predators. * Ranges of proportions assigned to individual predators are indicated for Diet 4.

Species	**Diet 1**	**Diet 2**	**Diet 3**	**Diet 4***
American lobster	0.00	0.00	0.00	0.00
American plaice	0.00	0.00	0.00	0.00
Atlantic argentine	0.00	0.00	0.00	0.00
Atlantic butterfish	0.00	0.00	0.00	0.00
Atlantic cod	0.00	0.00	0.12	0.00
Atlantic herring	0.10	0.20	0.00	0.05 - 0.20
Atlantic mackerel	0.00	0.00	0.00	0.00
Capelin	0.00	0.00	0.00	0.00
Gaspereau	0.00	0.00	0.00	0.00
Haddock	0.00	0.00	0.00	0.00
Longhorn sculpin	0.00	0.00	0.07	0.00
Marlin-spike grenadier	0.00	0.00	0.00	0.00
Monkfish	0.00	0.00	0.00	0.00
Northern sandlance	0.15	0.20	0.12	0.05 - 0.30
Northern shortfin squid	0.00	0.00	0.00	0.00
Northern shrimp	0.00	0.00	0.00	0.00
Ocean pout	0.00	0.00	0.00	0.00
Pollock	0.30	0.20	0.15	0.25 - 0.40
Red hake	0.00	0.00	0.00	0.00
Redfish	0.40	0.20	0.20	0.30 - 0.40
Sea raven	0.00	0.00	0.00	0.00
Silver hake	0.00	0.00	0.00	0.00
Smooth skate	0.00	0.00	0.00	0.00
Snake blenny	0.00	0.00	0.06	0.00
Snow crab	0.00	0.00	0.00	0.00
Toad crab	0.00	0.00	0.00	0.00
Thorny skate	0.00	0.00	0.08	0.00
White hake	0.05	0.20	0.08	0.05 - 0.10
Windowpane flounder	0.00	0.00	0.00	0.00
Winter flounder	0.00	0.00	0.00	0.00
Winter skate	0.00	0.00	0.00	0.00
Witch flounder	0.00	0.00	0.08	0.00
Yellowtail flounder	0.00	0.00	0.04	0.00

removal.

Valid estimates using the extended prey library could only be obtained through FS, which was performed on each of the 100 samples for each diet. MLE diet estimates were then obtained for each sample using the associated subset of selected prey. QFASA diet estimates for the same predators using a prey library limited to the exact species present in the true diet were obtained as an ideal baseline against which to compare the estimated MLE diet proportions for the true nonzero species; these QFASA estimates reflect an unrealistic best-case scenario.

For Diets 1-3, boxplots of the distributions of diet estimates across the 1500 pseudo-predators were obtained. For Diet 4, boxplots of the bias of estimates (calculated as the difference between the estimated and true dietary proportions) across the 1500 pseudo-predators were obtained.

Chi-squared distances between the true and estimated diet compositions were obtained for all pseudo-predators and boxplots of the distances grouped by diet were obtained. In the case of the basal prey library, these distances were obtained for BE and FS estimates and compared against QFASA estimates.

Match rates were calculated for each sample and each estimation method. Species were grouped as either true zero or true nonzero based on their proportion in the true diet, and the mean of the correct classification rate by method was calculated for each group from the 100 samples for each diet. For BE and FS, only those species omitted from the selected prey subset were considered to have been classified as zeroes. For QFASA, species estimated with a maximum dietary proportion of < 0.001 were considered to have been classified as zeroes, as in Goetsch et al. (2018). Species-specific rates were calculated as the percentage of iterations in which species were correctly classified.

3.3.3 Real-Life Example

In the real-life example, we obtained diet estimates for two samples of free-ranging adult northwest Atlantic grey seals from Dr. Shelley Lang (Research Scientist, Fisheries and Oceans Canada). Blubber samples from these seals were collected on Sable Island, Nova Scotia, Canada during the spring moulting seasons (May-June) of 1997 and 2004 as outlined in Beck et al. (2007). The 1997 sample was made up of 39 individual seals, 19 male and 20 female, and the 2004 sample was made up of 49 individuals, 25 male and 24 female. FA signatures comprised of 67 FAs were available for each seal, with the compositions reduced and renormalized during analysis according to the dietary subset of 28 FAs (Table 3.2). Values for fat content adjustment and calibration coefficients (CCs) are given in Tables 3.1 and 3.4 respectively.

Table 3.4: List of calibration coefficients (CCs) for the 28 dietary FAs.

FA	**CC**
16:2n-6	0.77
16:2n-4	1.50
16:3n-6	0.88
16:3n-4	0.69
16:4n-1	0.60
18:2n-6	2.02
18:2n-4	0.98
18:3n-6	1.10
18:3n-4	4.04
18:3n-3	2.25
18:3n-1	0.93
18:4n-3	0.95
18:4n-1	1.13
20:1n-11	3.44
20:1n-9	0.81
20:1n-7	0.75
20:2n-6	1.65
20:3n-6	1.67
20:4n-6	0.82
20:3n-3	2.20
20:4n-3	2.09
20:5n-3	0.65
21:5n-3	1.38
22:4n-6	3.29
22:5n-6	1.04
22:4n-3	2.74
22:5n-3	4.69
22:6n-3	1.10

Diet estimates using the basal prey library were obtained for BE, FS, and QFASA using both KL and Aitchison's distances. BE used a threshold of > 0.01 in the QFASA [AIT] diet estimates of all predators for a species to be considered for removal. Using the extended prey library, valid diet estimates were obtained using FS and for the sake of comparison, QFASA estimates using both KL and Aitchison's distances were also obtained. Boxplots of all diet estimates were obtained. Forward selection was carried out for each sample with the minimum prey subset size set at 10. No starting species were specified.

Chapter 4

Results

4.1 Simulation Studies

4.1.1 Case 1: # Species < # FAs

The quality of MLE estimates obtained using prey subsets selected by the FS and BE algorithms varied between diet groups (Figures 4.1, 4.2). Because we were able to obtain valid QFASA estimates in Case 1, QFASA [KL] estimates were included for comparison as this method is often used in practice and KL divergence is commonly used in QFASA applications.

For Diets 1, 2 and 4, FS estimates assigned smaller proportions to true zero (i.e., unconsumed) species than either QFASA or BE estimates for the same pseudo-predators, and were generally less variable (Figures 4.1, 4.2B). QFASA consistently assigned greater dietary proportions to true zero species than either selection method for these diets and underestimated the proportions of true nonzero (i.e., consumed) species. In the case of Diet 4, where pseudo-predators were assigned distinct dietary proportions, both FS and BE resulted in less biased estimates than QFASA (Figure 4.2B).

FS estimates were more variable and less accurate than the other methods for

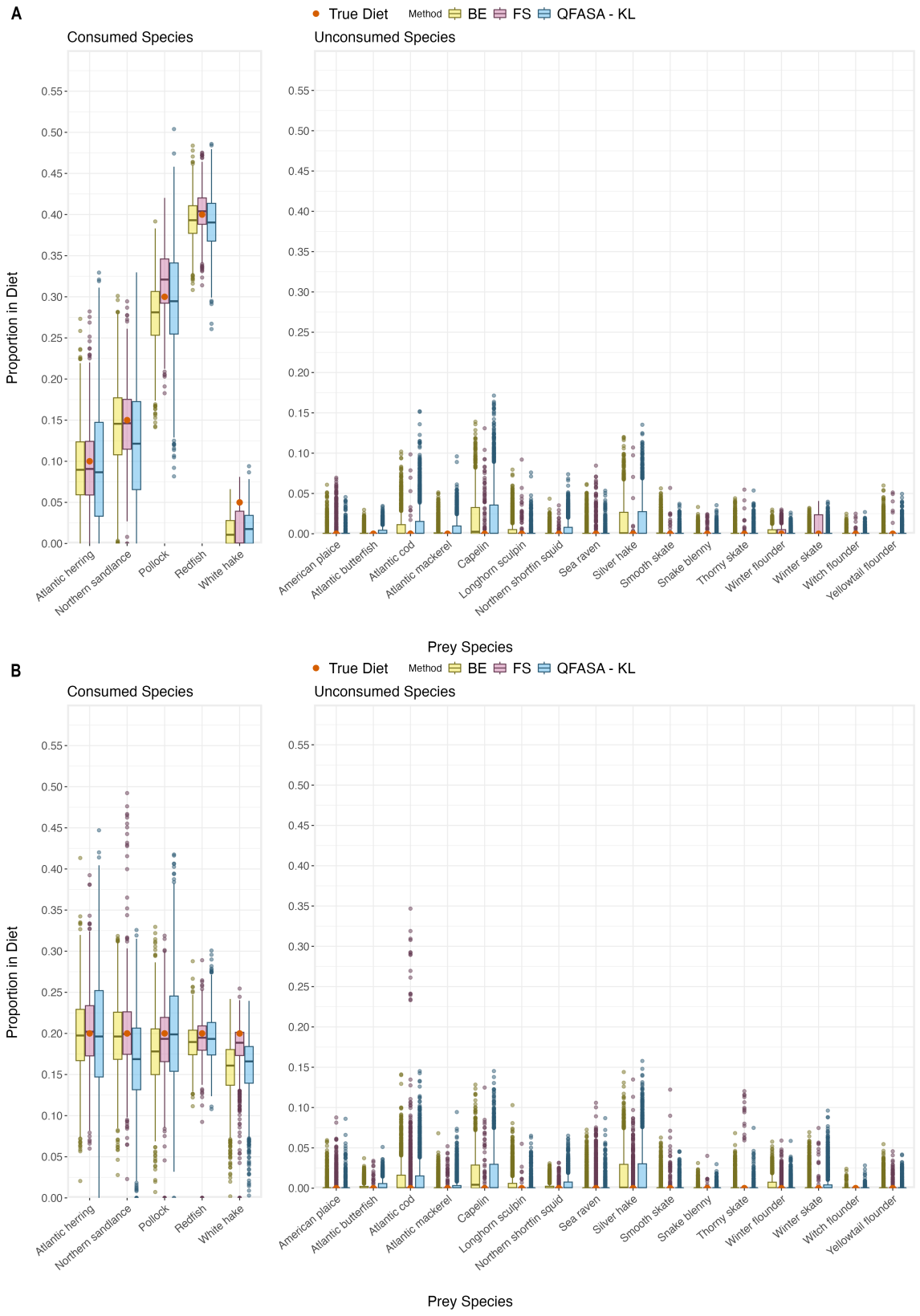


Figure 4.1: Diet estimates for 1500 individual pseudo-predators generated from Diet 1 (A) and Diet 2 (B).

Diet 3, where the number of species present in the true diet was increased (Figure 4.2A), with the maximum bias reaching as high as 0.46 in one predator's estimated proportion of Atlantic herring. For this diet, BE estimates were on average the most accurate for true nonzero species.

For all diets, chi-squared distances (Stewart, 2016) between estimates and true diets were similar for QFASA and BE estimates, while FS estimates resulted in notably smaller distances for Diets 1, 2, and 4, and greater distances for Diet 3 (Figure 4.3), which is in line with the conclusions of the diet estimate boxplots. While FS chi-squared distances were greater in the case of Diet 3, it maintained the best match rates for species classification and therefore, may be preferable when correctly excluding species is the objective (see discussion below).

There were no instances of true nonzero species being excluded by either selection method in any of the 400 total samples. FS resulted in significantly smaller subsets of selected prey on average than BE for all diets. The mean rate for the correct classification of true zero species ranged between diets from 70.5 - 92.3% for FS and 7.3 - 22.6% for BE (Table 4.1).

At the species-specific level, match rates varied widely among true zero species. For Diet 2, BE failed in every sample to exclude three species (Atlantic cod, capelin, and silver hake), while witch flounder was correctly excluded 84% of the time and snake blenny 73% of the time. FS was more consistent in this regard; across all diets, 40/58 species-specific match rates were $\geq$90% for this method. Its lowest match rate was 36% for sea raven in Diet 3 (Table 4.2).

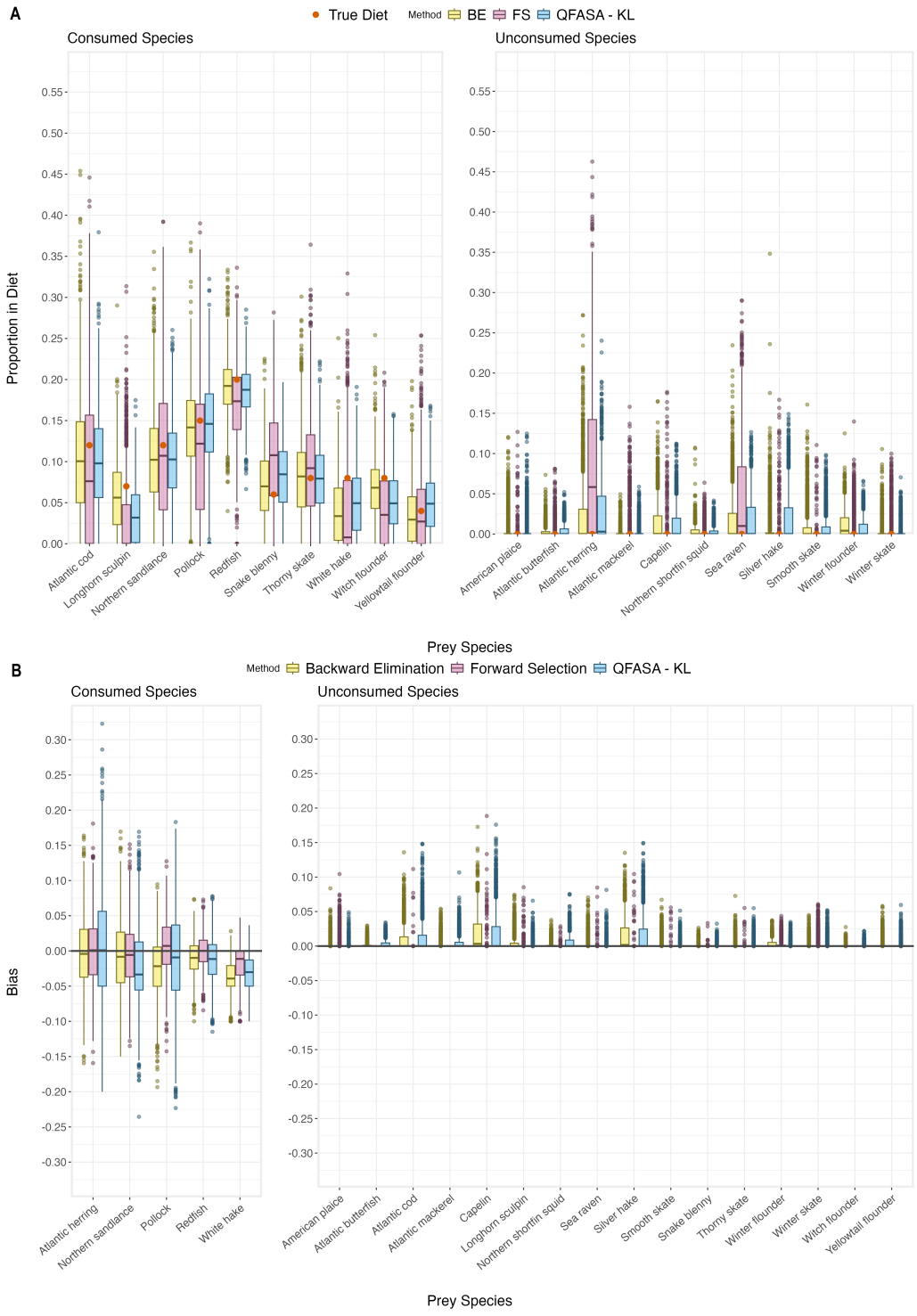


Figure 4.2: Diet estimates for 1500 individual pseudo-predators generated from Diet 3 (A). Biases (estimated proportion - true proportion) on diet estimates for 1500 individual pseudo-predators generated from Diet 4 (B).

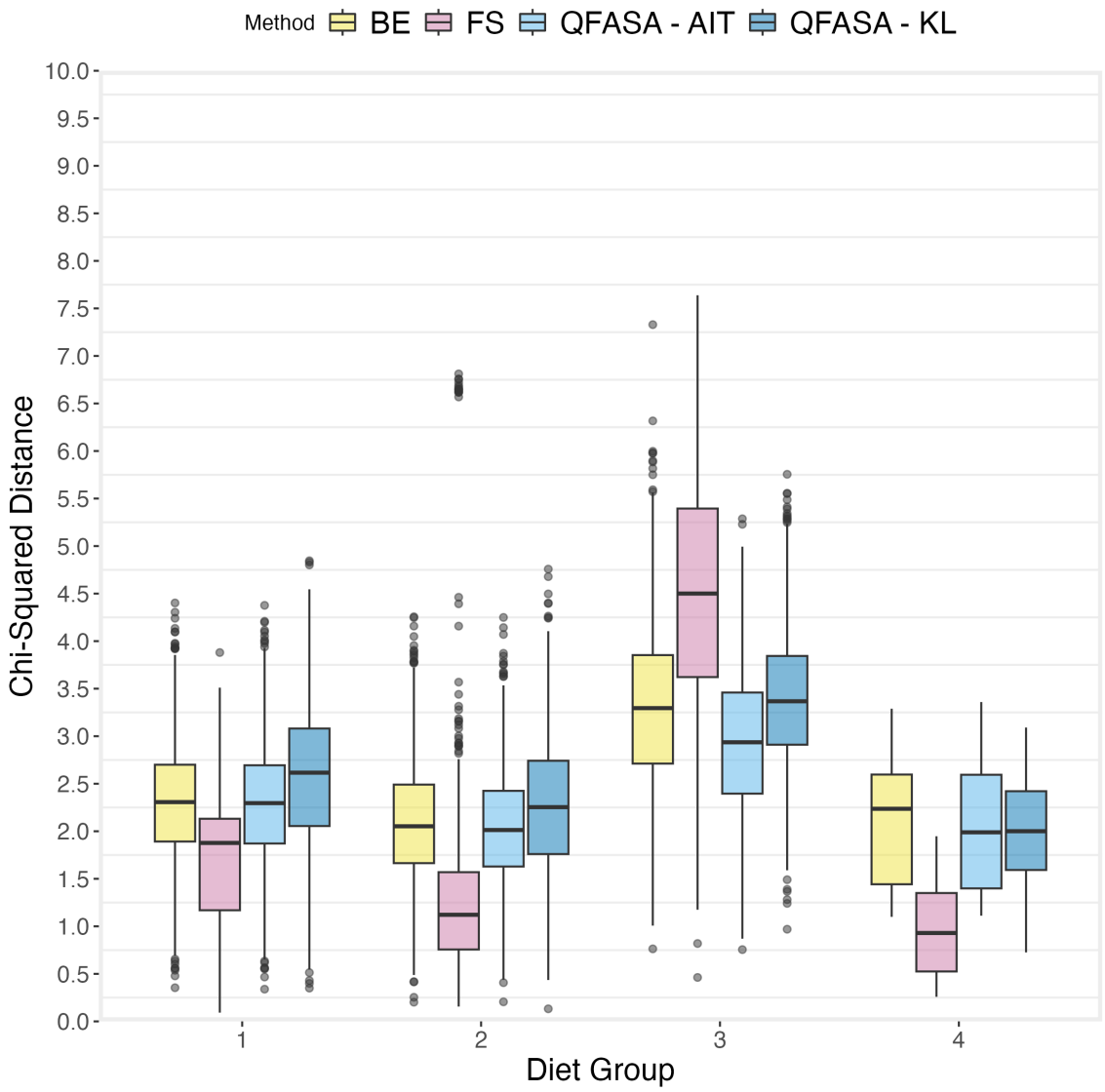


Figure 4.3: Chi-squared distances between true and estimated diet compositions for 1500 pseudo-predators generated from each diet.

Table 4.1: Match rates (mean and SD as percentages) of the correct classification of true zero and true nonzero species by method for each diet.

	True Zero		**True Nonzero**	
Method	**Correct**	**Incorrect**	**Incorrect**	**Correct**
Diet 1				
BE	22.56 (7.69)	77.44 (7.69)	0 (0)	100 (0)
FS	90.75 (4.12)	9.25 (4.12)	0 (0)	100 (0)
QFASA - AIT	5.38 (5.26)	94.62 (5.26)	0 (0)	100 (0)
QFASA - KL	13.31 (8.41)	86.69 (8.41)	0 (0)	100 (0)
Diet 2				
BE	21.00 (7.25)	79.00 (7.25)	0 (0)	100 (0)
FS	87.62 (5.89)	12.38 (5.89)	0 (0)	100 (0)
QFASA - AIT	5.31 (5.05)	94.69 (5.05)	0 (0)	100 (0)
QFASA - KL	17.38 (9.55)	82.62 (9.55)	0 (0)	100 (0)
Diet 3				
BE	7.27 (7.42)	92.73 (7.42)	0 (0)	100 (0)
FS	70.45 (11.22)	29.55 (11.22)	0 (0)	100 (0)
QFASA - AIT	2.36 (4.40)	97.64 (4.40)	0 (0)	100 (0)
QFASA - KL	1.45 (3.82)	98.55 (3.82)	0 (0)	100 (0)
Diet 4				
BE	21.81 (8.26)	78.19 (8.26)	0 (0)	100 (0)
FS	92.31 (3.65)	7.69 (3.65)	0 (0)	100 (0)
QFASA - AIT	5.38 (5.26)	94.62 (5.26)	0 (0)	100 (0)
QFASA - KL	13.12 (7.72)	86.88 (7.72)	0 (0)	100 (0)

Table 4.2: Species-specific match rates of the correct classification of true zero species as a percentage of 100 iterations per diet. Dashes indicate rates that do not exist as the species is not a true zero in that diet.

	Diet 1		**Diet 2**		**Diet 3**		**Diet 4**	
Species	**BE**	**FS**	**BE**	**FS**	**BE**	**FS**	**BE**	**FS**
American plaice	43	80	34	79	7	91	37	79
Atlantic butterfish	29	100	19	97	9	81	21	100
Atlantic cod	1	99	0	47	-	-	1	99
Atlantic mackerel	8	100	18	97	33	77	12	100
Capelin	0	95	0	96	0	88	0	96
Longhorn sculpin	6	97	4	99	-	-	8	96
Northern shortfin squid	23	98	19	92	6	73	20	98
Sea raven	10	94	9	68	2	36	10	98
Silver hake	0	99	0	76	3	90	0	98
Smooth skate	6	99	13	95	1	91	13	99
Snake blenny	67	98	73	99	-	-	71	99
Thorny skate	27	99	23	99	-	-	26	99
Winter flounder	10	50	3	67	0	73	8	48
Winter skate	30	46	13	98	19	75	38	68
Witch flounder	79	99	84	100	-	-	73	100
Yellowtail flounder	22	99	24	93	-	-	11	100

4.1.2 Case 2: # Species $\geq$ # FAs

In this case, valid estimates could only be obtained using FS. For the sake of comparison with the best-case scenario of estimates using the perfectly selected prey library, QFASA [AIT] estimates were obtained using a prey library artificially reduced to include only the species in the true diet; these estimates represent an unrealistic best-case scenario.

As in the previous case, performance varied between true diet groups. In those diets comprised of fewer true nonzero species (Diets 1, 2, and 4), median estimated proportions fell within 0.05 of the true assigned proportion for all species (Figures 4.4, 4.5B). However, for Diet 3, median estimated proportions were up to 0.1 from the true assigned proportion, with the maximum bias for a single individual estimate reaching up to 0.55 (Figure 4.5A).

For Diets 1, 2 and 4, FS estimates for most true nonzero species were distributed very similarly to QFASA estimates obtained using Aitchison's distance and exactly the correct subset of true prey species, an unrealistic best-case scenario. In exception, white hake proportions tended to be underestimated and more variable using FS, though medians remained within 0.05 of the true diet. For this purpose, Aitchison's distance was chosen for the comparison to illustrate the best-case estimates due to the relationship outlined in Chapter 3.

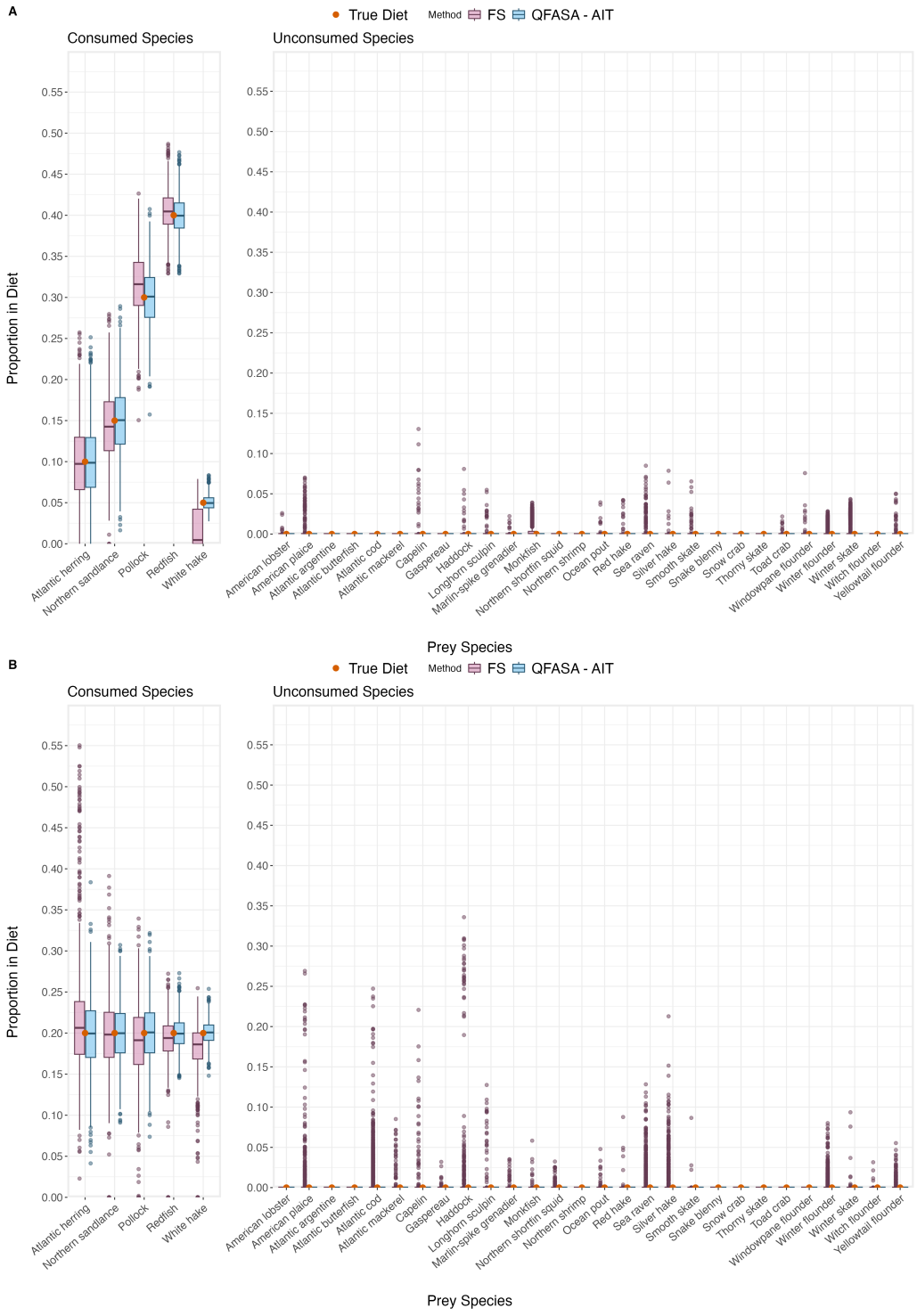


Figure 4.4: Diet estimates for 1500 individual pseudo-predators generated from Diet 1 (A) and Diet 2 (B) following FS on the extended prey library compared against QFASA estimates using the exact subset of species included in the true diet.

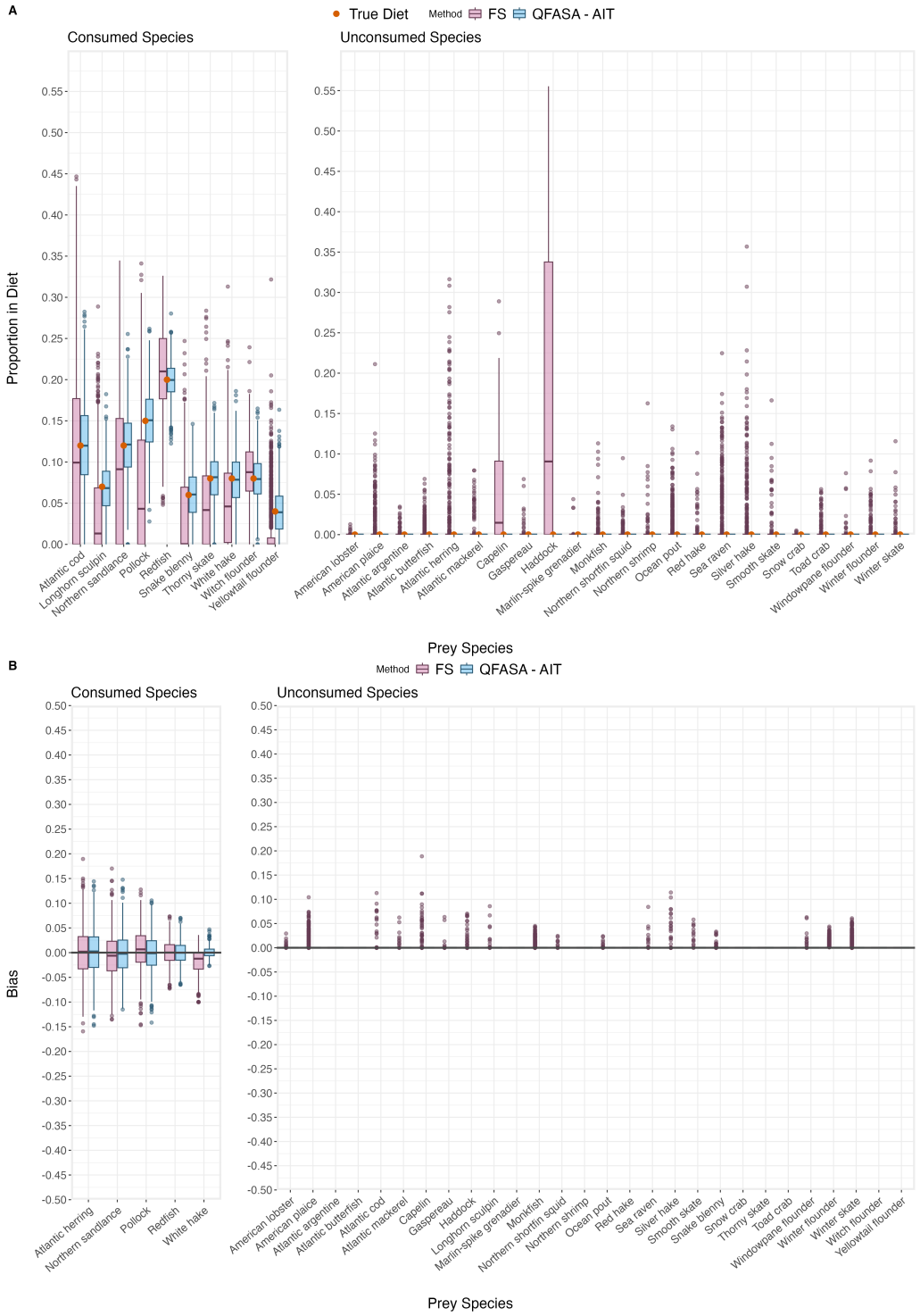


Figure 4.5: Diet estimates for 1500 individual pseudo-predators generated from Diet 3 following FS on the extended prey library (A). Biases (estimated proportion - true proportion) on diet estimates for 1500 individual pseudo-predators generated from Diet 4 following FS on the extended prey library (B). Both are compared against QFASA estimates using the exact subset of species included in the true diet.

There were no instances of true nonzero species being incorrectly classified by FS in any of the 400 total iterations. The mean rate for the correct classification of true zero species ranged from 85% (Diet 3) to 95% (Diet 4). This rate varied most significantly for Diet 3 (SD = 5.84 %), which displayed a less consistent choice in selected prey subset between samples than other diets (Table 4.3). The haddock introduced in the extended prey library was incorrectly selected in at least some iterations of each diet, and in all iterations of Diet 3. Capelin was also poorly classified, having been selected in 93% of iterations for Diet 3. Half of all true zero species were excluded with accuracy greater than 90% across all diets (Table 4.4).

Table 4.3: Match rates (mean and SD as percentages) of the correct classification of true zero and true nonzero species by FS for each diet using the extended prey library.

	True Zero		**True Nonzero**	
Diet	**Correct**	**Incorrect**	**Incorrect**	**Correct**
1	95.21 (2.23)	4.79 (2.23)	0 (0)	100 (0)
2	92.5 (3.61)	7.5 (3.61)	0 (0)	100 (0)
3	84.96 (5.84)	15.04 (5.84)	0 (0)	100 (0)
4	95.61 (1.67)	4.39 (1.67)	0 (0)	100 (0)

Table 4.4: Species-specific match rates of the correct classification of true zero species as a percentage of 100 iterations per diet using the extended prey library.

Species	**Diet 1**	**Diet 2**	**Diet 3**	**Diet 4**
American lobster	99	100	98	99
American plaice	92	81	88	85
Atlantic argentine	100	100	96	100
Atlantic butterfish	100	100	86	100
Atlantic mackerel	100	97	94	99
Capelin	98	96	7	96
Gaspereau	100	99	97	99
Haddock	99	83	0	97
Marlin-spike grenadier	99	98	99	100
Monkfish	58	99	91	84
Northern shortfin squid	100	96	95	99
Northern shrimp	100	100	95	100
Ocean pout	99	97	82	99
Red hake	99	99	94	100
Sea raven	94	65	87	99
Silver hake	99	73	87	98
Smooth skate	98	99	94	99
Snow crab	100	100	99	100
Toad crab	99	100	92	100
Windowpane flounder	99	100	98	99
Winter flounder	63	70	95	52
Winter skate	75	99	96	78

4.2 Real-Life Example

4.2.1 Diet Estimates

4.2.1.1 Case 1: # Species < # FAs

Figure 4.6 shows diet estimates for the 1997 sample using the basal prey library (Case 1). Redfish had the highest median estimated proportion for all methods, though the median with QFASA [KL] was > 0.15 less than the other methods. When comparing median estimates by method, the other species with noticeable differences are Atlantic herring (0.05 with FS, 0.01 with QFASA [KL], and 0 with BE and QFASA [AIT]), longhorn sculpin (0.29 with FS, 0.27 with BE, 0.35 with QFASA [AIT], and 0 with QFASA [KL]), and white hake (0.05 with QFASA [KL] and 0 for all others).

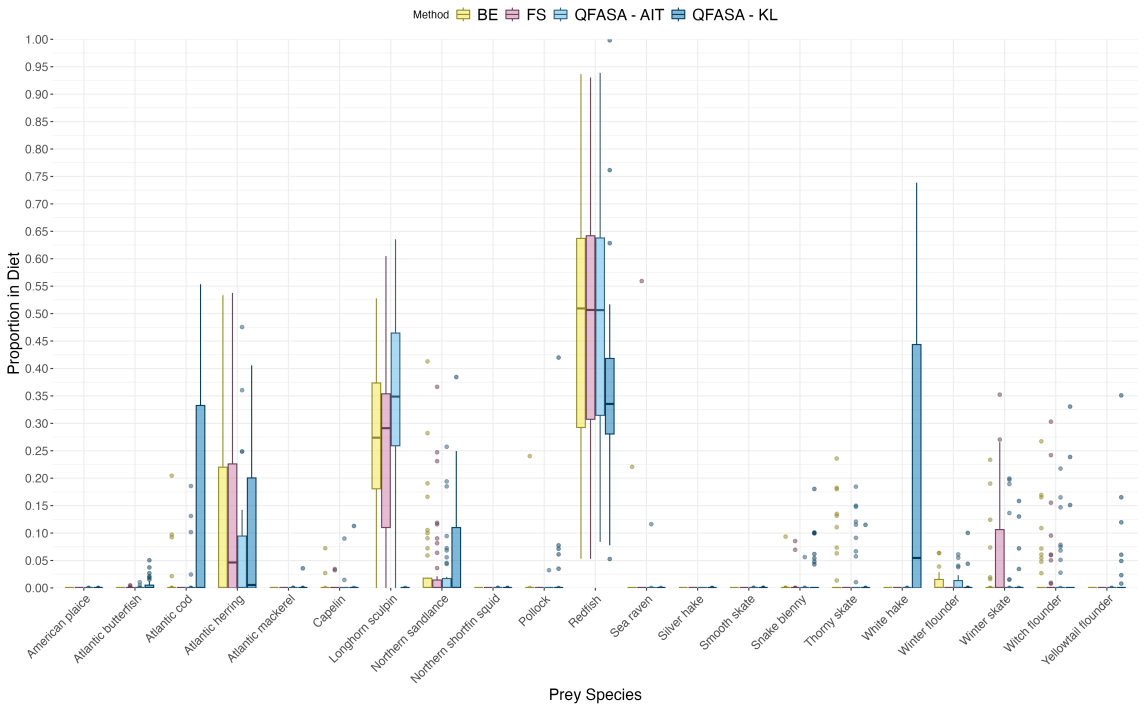


Figure 4.6: Diet estimates for 39 adult grey seals sampled in 1997 using the basal prey library with a minimum model size of 10. Backward elimination (BE) and forward selection (FS) estimates are compared with QFASA estimates obtained using both Aitchison's distance (AIT) and Kullback-Leibler divergence (KL).

Figure 4.7 shows diet estimates for the 2004 sample using the basal prey library (Case 1). Redfish again had the highest median estimated proportion for all methods (0.60 with FS, 0.61 with BE, 0.54 with QFASA [AIT], and 0.41 with QFASA [KL]). Median estimates for longhorn sculpin were 0.05 and 0.18 for BE and QFASA [AIT] respectively and 0 for FS and QFASA [KL]. FS and QFASA [KL] estimates were more similar to each other for northern sandlance and white hake, though the median estimate for white hake was much higher for QFASA [KL] at 0.17 compared with < 0.01 for the other methods. The median estimate for winter skate was 0.13 for FS and < 0.01 for the other methods.

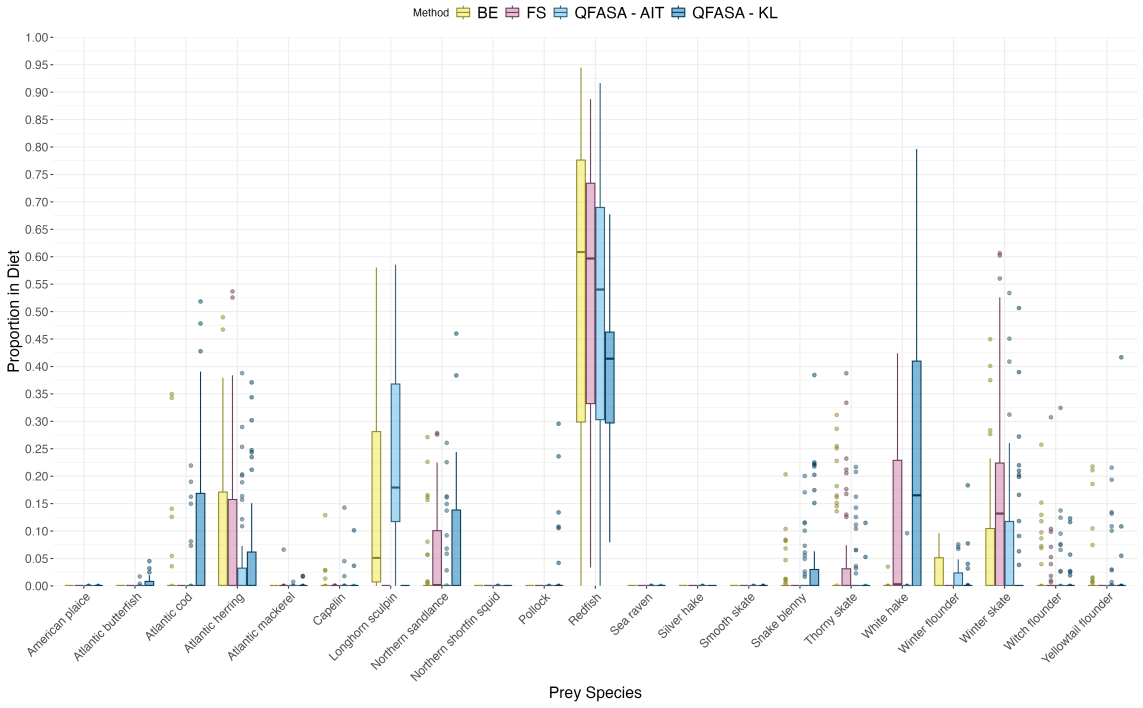


Figure 4.7: Diet estimates for 49 adult grey seals sampled in 2004 using the basal prey library with a minimum model size of 10. Backward elimination (BE) and forward selection (FS) estimates are compared with QFASA estimates obtained using both Aitchison's distance (AIT) and Kullback-Leibler divergence (KL).

4.2.1.2 Case 2: # Species $\geq$ # FAs

While QFASA estimates may not be unique in the case of more species than fatty acids, they are included for the sake of comparison in Figures 4.8 and 4.9, which show diet estimates for the seals sampled in 1997 and 2004 respectively following forward selection on the extended prey library.

For the 1997 sample, FS estimates seem fairly similar to QFASA [AIT] estimates (Figure 4.8). Redfish remains the species with the highest estimated proportions. Comparing median diet estimates by method, the species with differences are Atlantic herring (0.01 with QFASA [KL] and < 0.01 with FS and QFASA [AIT]), longhorn sculpin (0.20 with FS, 0.33 with QFASA [AIT], and 0 with QFASA [KL]), and white hake (0.02 with QFASA [KL] and 0 with FS and QFASA [AIT]). While differences in median estimates for Atlantic cod and white hake may be insignificant, differences in the variability of estimates for these species are evident in the boxplot. Proportions up to 0.47 of the newly added species Atlantic argentine were estimated for some individual predators with QFASA [KL], which differed greatly from the QFASA [AIT] estimates and FS where it was excluded entirely.

For the 2004 sample using the extended prey library, all methods again estimated redfish to be the species consumed in the highest proportions (Figure 4.9). FS estimates differed more from QFASA [AIT] estimates in this setting than in the others. In this case, many species estimated to be in the diets of several individual predators by both methods of QFASA were excluded by FS. These species were Atlantic argentine, Atlantic cod, monkfish, snake blenny, winter flounder, winter skate, witch flounder and yellowtail flounder, and while they all had median estimates < 0.000001, their mean estimates with QFASA ranged from 0.01 up to 0.10. Species with median FS estimates $\geq$0.01 were redfish (0.55), white hake (0.27), marlin-spike grenadier (0.01) and northern sandlance (0.01).

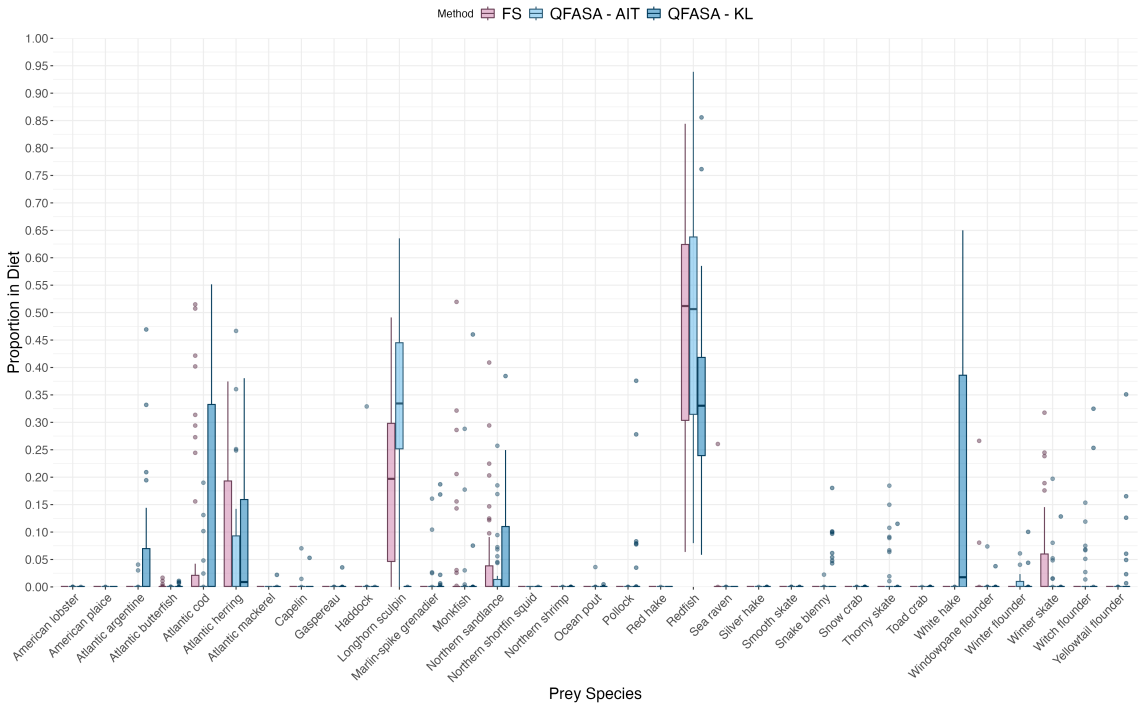


Figure 4.8: Diet estimates for 39 adult grey seals sampled in 1997 using the extended prey library with a minimum model size of 10. Forward selection (FS) estimates are compared with QFASA estimates obtained using both Aitchison's distance (AIT) and Kullback-Leibler divergence (KL).

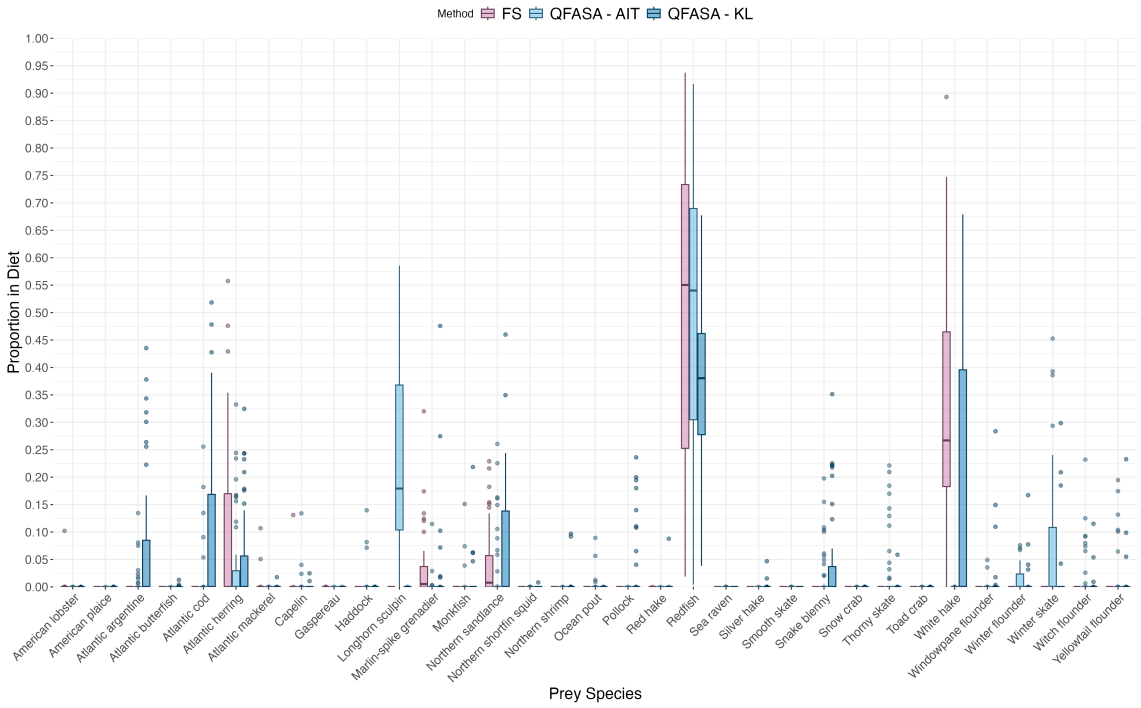


Figure 4.9: Diet estimates for 49 adult grey seals sampled in 2004 using the extended prey library with a minimum model size of 10. Forward selection (FS) estimates are compared with QFASA estimates obtained using both Aitchison's distance (AIT) and Kullback-Leibler divergence (KL).

4.2.2 Species Selection

4.2.2.1 Backward Elimination

For Case 1, the order in which species were removed using backward elimination is outlined in Table 4.5. For both samples, a total of eight species were removed. Six of the species removed were common between samples, with the differences being yellowtail flounder and white hake removed for the 1997 sample and sea raven and pollock removed for the 2004 sample.

There were some notable removals of species present in the QFASA [KL] diet estimates of some predators. For the 1997 sample, these were yellowtail flounder with a maximum estimate of 0.35, white hake with a median estimate of 0.05, Atlantic mackerel with a maximum estimate of 0.04, and Atlantic butterfish with a maximum estimate of 0.05. For the 2004 sample, they were Atlantic mackerel with a maximum estimate of 0.02, Atlantic butterfish with a maximum estimate of 0.05, and pollock with a maximum estimate of 0.30. With the exception of white hake, the median QFASA [KL] estimates for these species were > 0.01.

Table 4.5: Removal order of species added using backward elimination with the basal prey library (Case 1).

Removal Order	**1997**	**2004**
1	Yellowtail flounder	Sea raven
2	White hake	Northern shortfin squid
3	Northern shortfin squid	Atlantic mackerel
4	American plaice	Atlantic butterfish
5	Silver hake	Smooth skate
6	Atlantic mackerel	Pollock
7	Atlantic butterfish	Silver hake
8	Smooth skate	American plaice

4.2.2.2 Forward Selection

Tables 4.6 and 4.7 show the selection order of species using forward selection in Cases 1 and 2 respectively. The first two species ordered "1" are the pair selected

together in the first step of the algorithm.

In Case 1 (Table 4.6), the choice of starting species differed between samples. The pair of longhorn sculpin and redfish selected for the 1997 sample was the third best candidate for the 2004 sample, with an AIC 13.6% higher than the best candidate. The pair of redfish and white hake selected for the 2004 sample was also the third best candidate for the 1997 sample, with an AIC 5.3% higher than the best candidate. Longhorn sculpin was not later selected for the 2004 sample despite appearing in its QFASA [AIT] and BE estimates, and white hake was not later selected for the 1997 sample despite appearing in its QFASA [KL] estimates. Other differences between samples include the selection of sea raven, smooth skate and winter flounder for the 1997 sample, and Atlantic mackerel, silver hake and thorny skate for the 2004 sample.

Table 4.6: Selection order of species added using forward selection with the basal prey library (Case 1). Species marked with an asterisk are those that would have been included without a minimum model size imposed.

Selection Order	**1997**	**2004**
1	Longhorn sculpin*	Redfish*
1	Redfish*	White hake*
3	Atlantic butterfish	Atlantic herring*
4	Witch flounder	Thorny skate
5	Snake blenny	Capelin
6	Capelin	Winter skate
7	Smooth skate	Northern sandlance
8	Witch flounder	Silver hake
9	Sea raven	Atlantic mackerel
10	Winter flounder	Witch flounder

In Case 2 (Table 4.7), the choice of starting species again differed between samples but both remained the same as in Case 1. Other differences between samples include the selection of Atlantic butterfish, Atlantic cod, sea raven, monkfish and windowpane flounder for the 1997 sample, and American lobster, Atlantic mackerel, capelin, gaspereau and red hake for the 2004 sample.

Without the implementation of a minimum model size, FS included very few species for each case and sample, ranging from just 2 (Case 1, 1997) to 4 (Case 2, 2004).

Table 4.7: Selection order of species added using forward selection with the extended prey library (Case 2). Species marked with an asterisk are those that would have been included without a minimum model size imposed.

Selection Order	**1997**	**2004**
1	Longhorn sculpin*	Redfish*
1	Redfish*	White hake*
3	Atlantic herring*	Atlantic herring*
4	Northern sandlance	Atlantic mackerel*
5	Winter flounder	American lobster
6	Sea raven	Red hake
7	Monkfish	Marlin-spike grenadier
8	Atlantic butterfish	Capelin
9	Atlantic cod	Gaspereau
10	Windowpane flounder	Northern sandlance

4.2.2.3 Candidate Models

In some cases, there were several candidate models with similar AIC values at a given step. Table 4.8 shows, for the first three iterations of forward selection, the range of AIC values (highest - lowest) associated with all possible candidate models at that step, and the number of candidate models whose AIC is within 10% of the lowest.

The sample from 1997 had less range in AIC values and more similarly fit candidate models with both prey libraries (Case 1 and Case 2). In the first step, the selection of the best pair of starting species, there were four and six candidates within 10% of each other for the respective cases. In Case 1, the AIC of the second best candidate was only 0.44% higher than the best in step 1, and fell to only 0.07% in step 2 where six candidates (Atlantic butterfish, thorny skate, northern sandlance, smooth skate, winter skate and Atlantic herring) had AIC values within 1% of each

other. There were more obvious differences between candidates for the sample from 2004. In Case 1, the second best choice was 8.7% higher than the best in step 1 and 7.8% higher in step 2.

In both cases and for both samples, the top candidates for starting species all featured redfish and moreover, all pairs including redfish had lower AIC values than any pair without it.

Table 4.8: Range of AIC values for the first three iterations of forward selection for seals sampled in 1997 and 2004 using basal (Case 1) and extended (Case 2) prey libraries. The number of candidate models whose AIC value is within 10% of the best is shown as a fraction of the total number of candidate models for that iteration.

		Case 1		**Case 2**	
		1997	**2004**	**1997**	**2004**
Step 1	**AIC Range**	2794.10	3284.33	2862.43	3725.7
	# of candidates within 10% of best	4/210	2/210	6/528	2/528
Step 2	**AIC Range**	45.78	275.18	61.17	290.84
	# of candidates within 10% of best	19/19	2/19	31/31	2/31
Step 3	**AIC Range**	66.94	190.85	94.21	223.06
	# of candidates within 10% of best	18/18	12/18	30/30	14/30

4.2.2.4 Selection Differences Between Cases

Forward selection estimates for each sample also changed from Case 1 to Case 2 (Figure 4.10). For the 1997 sample, the difference in selected species were capelin, smooth skate, winter flounder and witch flounder in Case 1 being replaced with Atlantic butterfish, Atlantic cod, monkfish and windowpane flounder in Case 2. Of the four included in Case 2, two were exclusive to the extended prey library and two were not.

Estimates for this sample did not change much on average, with the largest change being a reduction in the estimated proportion of longhorn sculpin from Case 1 to Case 2 (mean difference -0.07), but changes for some individual predators were

more significant. 13 of 39 individual predators had changes greater than 0.2 in the estimated dietary proportion of at least one species between Case 1 and Case 2. At the individual level, these changes saw proportions of Atlantic herring and winter skate moving to monkfish and sea raven moving to Atlantic cod, while longhorn sculpin and witch flounder moved to both Atlantic cod and monkfish.

For the 2004 sample, the difference in selected species were thorny skate, winter skate, silver hake and witch flounder in Case 1 being replaced with American lobster, red hake, marlin-spike grenadier and gaspereau in Case 2. All four of these were among the new species introduced in the extended prey library.

Estimates for this sample saw some larger changes on average, most notably an increase in white hake (mean difference 0.21) and corresponding decrease in winter skate (mean difference -0.15). At the individual level, 27 of 49 predators had changes greater than 0.2 in the estimated dietary proportion of at least one species between Case 1 and Case 2. For these individuals, proportions of thorny skate went to marlin-spike grenadier and white hake, proportions of winter skate went to white hake and Atlantic herring, and proportions of witch flounder went to white hake.

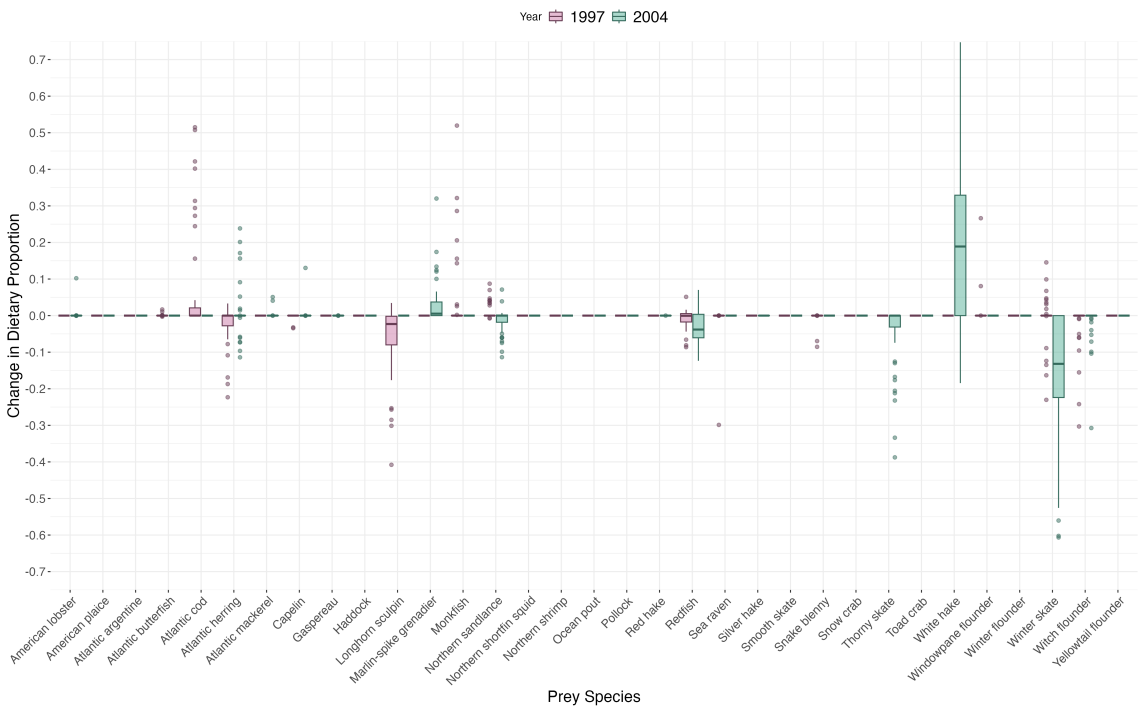


Figure 4.10: Change in estimated dietary proportions obtained using forward selection with the basal (Case 1) and extended (Case 2) prey libraries. Values reflect the change from Case 1 to Case 2.

4.3 Analysis of Prey Data

Interpreting prey selection outcomes requires some understanding of the underlying structure of the prey database and the relationships between species. The exploratory analyses performed in this section are available in the QFASA R package, and may be helpful in identifying potentially challenging species and subsets of species specific to the prey database.

Dendrograms displaying the results of a hierarchical cluster analysis on the mean prey FA signatures, generated by the prey.cluster function in the QFASA R package, show some differences between species groupings for chosen distance measures and species subsets (Figures 4.11, 4.12). When using the basal prey library, yellowtail flounder was clustered differently by KL than Aitchison's distance. However, with the addition of species in the extended prey library, clustering became more consistent between the two methods. Redfish, which was found to be the dominant prey in the real-life examples, remains distant from other species across both distance measures and prey libraries.

Closely-clustered species may be more difficult to differentiate from one another for the purposes of selection or estimation. Among the most notable clusters are the pair of longhorn sculpin and sea raven, and the group of Atlantic herring, capelin and northern sandlance. In simulation studies, when longhorn sculpin was included in the true diet, it was sometimes underrepresented in FS estimates while sea raven was overrepresented (Figure 4.2A). For the same diet group, northern sandlance was underestimated while Atlantic herring was overestimated.

Further analyses carried out with the prey.on.prey function in the QFASA R package provide some insight into the behaviour of individual species and their interactions in diet estimation. Using this method, diet estimates are obtained for each individual prey as though it were itself a predator eating from the remainder of the prey database (Bromaghin, 2017). We obtained these estimates using both

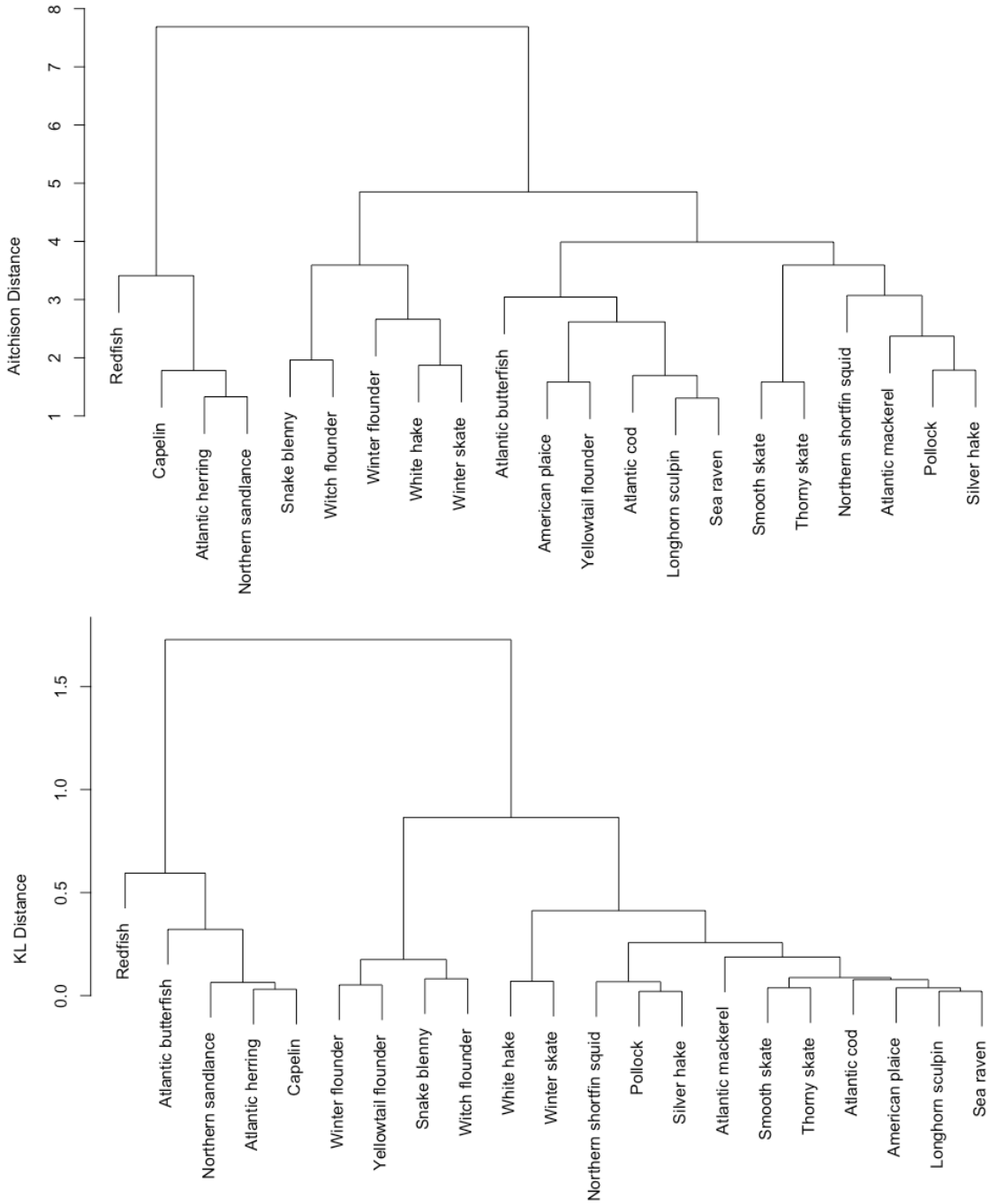


Figure 4.11: Clustering of species within the basal prey library as determined by Aitchison's (top panel) and KL (bottom panel) distances between mean FA signatures.

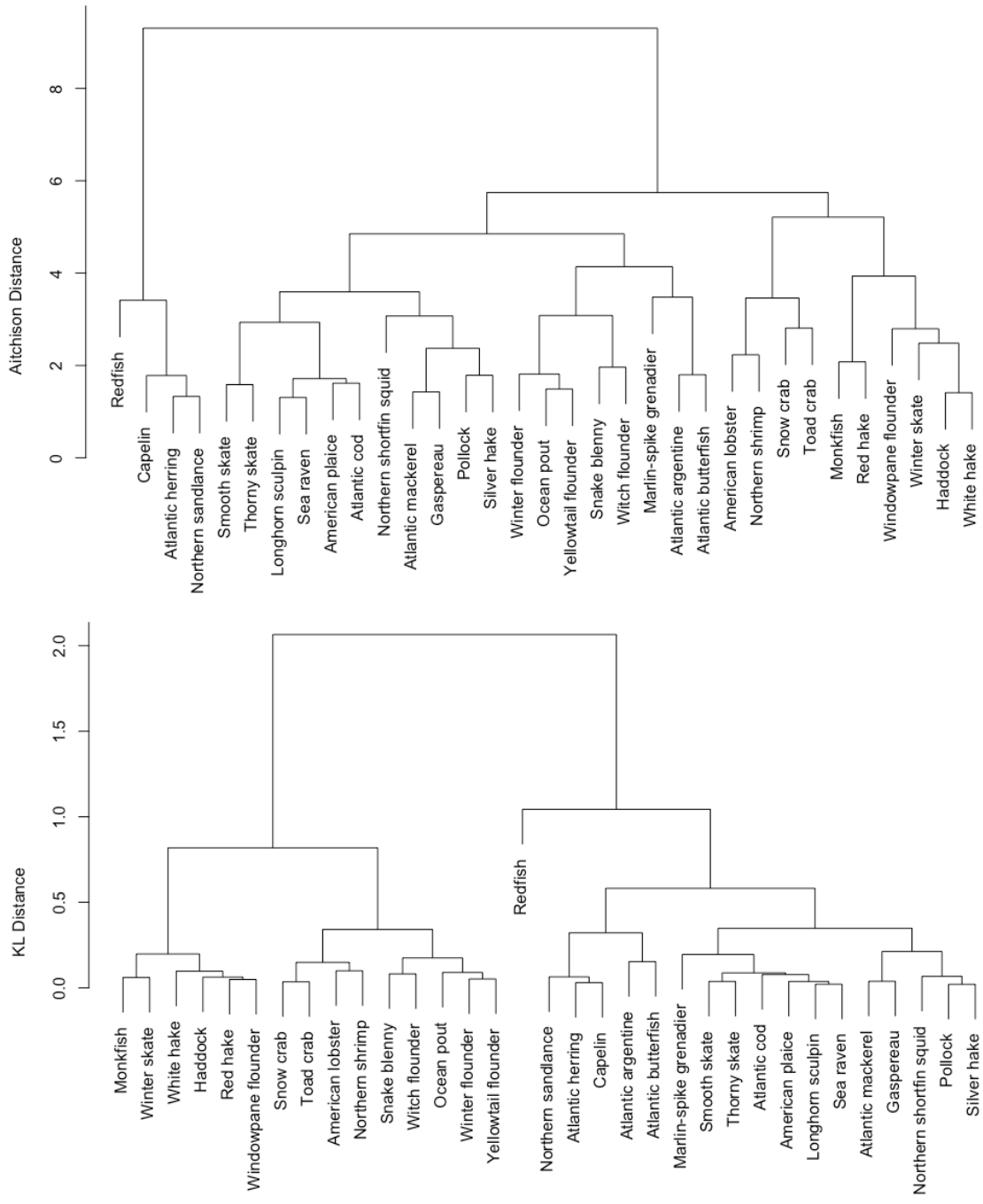


Figure 4.12: Clustering of species within the extended prey library as determined by Aitchison's and KL distances between mean FA signatures.

KL and Aitchison's distance. If conditions were perfect for estimation, each individual would be found to be eating its own species exclusively (i.e., an individual from species k should have an estimated dietary proportion of 1 for species k and 0 elsewhere). In practice, these values can deviate greatly from that ideal, thus creating more challenging conditions for accurate estimation and prey selection.

Mean diet estimates for each species in the extended prey library, obtained through prey-on-prey analysis, are shown in Figure 4.13. Each row sums to 1 and represents the mean estimated diet composition of the corresponding y-axis species. Values along the diagonal measure the mean estimated dietary proportion of each prey species in its own diet. Species with diagonal values nearer to 1 (e.g., redfish, snow crab and witch flounder) might be expected to be more easily identifiable among prey candidates, while those with diagonal values nearer to 0 (e.g., haddock, silver hake, white hake and yellowtail flounder) may be more difficult. This is reflected in the Case 2 simulation study estimates for diet group 3 (Figure 4.5A), where median redfish, snow crab and witch flounder estimates are within 1% of the true diet while median haddock, white hake and yellowtail flounder estimates are much less accurate.

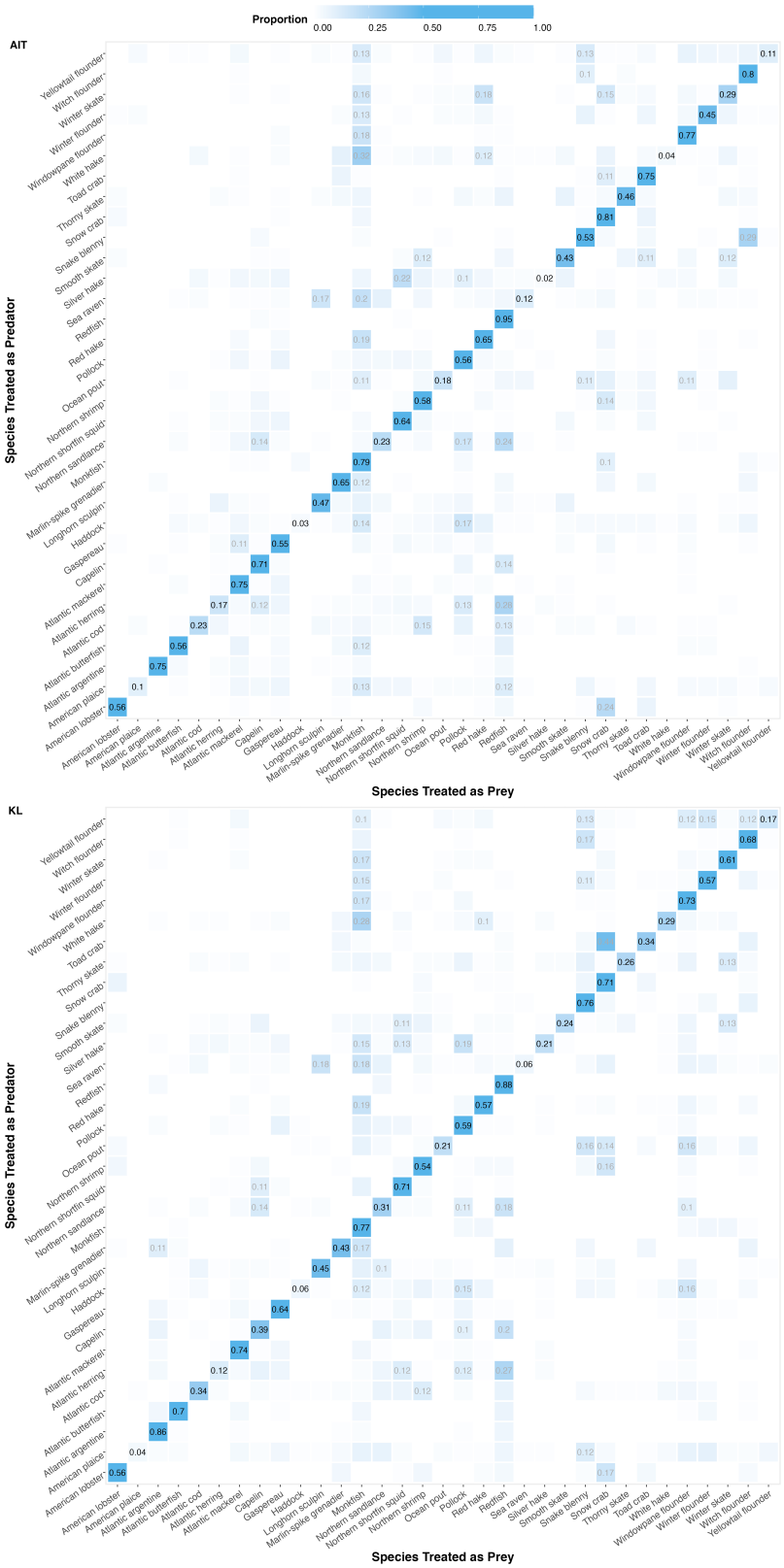


Figure 4.13: Mean prey-on-prey diet estimates for each species in the extended prey library using KL and Aitchison's distance.

The 13 least identifiable species in the extended prey library are summarized in Table 4.9. While both distance measures struggled to correctly assign several species, there were some notable differences between the two. KL did a much better job of correctly assigning silver hake and white hake at rates of 0.21 and 0.29 respectively, compared to just 0.02 and 0.04 using Aitchison's distance. This is notable given the diet estimates for the real-life examples in Section 4.2, where QFASA [KL] estimates for white hake were much higher than other methods for the 2004 sample in Case 1 and the 1997 sample in both cases. Haddock was the least correctly assigned species as averaged between the two distance measures, with rates of just 0.03 (AIT) and 0.06 (KL), and in Case 2 simulation studies, it was the only species with a match rate of 0 in any of the four diet groups (Table 4.4).

Table 4.9: Mean estimated dietary proportion assigned to select species by individuals of the same species using prey-on-prey analysis.

Species	**AIT**	**KL**
American plaice	0.10	0.04
Atlantic cod	0.23	0.34
Atlantic herring	0.17	0.12
Haddock	0.03	0.06
Northern sandlance	0.23	0.31
Ocean pout	0.18	0.21
Sea raven	0.12	0.06
Silver hake	0.02	0.21
Smooth skate	0.43	0.24
Thorny skate	0.46	0.26
White hake	0.04	0.29
Winter skate	0.29	0.61
Yellowtail flounder	0.11	0.12

These analyses also reveal the propensity for some species to be incorrectly attributed to various other species in diet estimates. For example, diet estimates for Atlantic herring using KL distance show a mean proportion of 0.14 assigned to the correct species, while the proportion attributed to redfish is twice as great at 0.28.

We can further determine which species tend to be generally over- or under-

represented in the prey-on-prey estimates across all species in the prey library. Consider that each column in Figure 4.13 represents the estimated proportions of the corresponding x-axis species in the diets of the species along the y-axis. If the perfect diet estimate for each species is a value of 1 for itself (on the diagonal) and 0 elsewhere, then under ideal conditions each column would sum to 1, although they are not constrained to. In fact, the only sum constraint on the columns is that together, their cumulative sum is $k \cdot 1 = k$, where k is the number of species and 1 is by definition the sum of a diet composition. Then for an x-axis species that is correctly represented in the diet estimates of the prey library, the sum of its corresponding column should be 1 and its proportion in the sum of all k diet compositions should be $\frac{1}{k}$.

Figure 4.14 shows the proportional representation of each species in the extended prey library in the diet estimates, with a value of 0 indicating that a species has been assigned the expected overall proportion of $\frac{1}{33}$. Some species (e.g., American plaice, white hake and yellowtail flounder) were under-represented, and others (e.g., monkfish, redfish and snow crab) were over-represented. In simulation studies, there was a notable downward bias in FS estimates for white hake compared to estimates for other nonzero species.

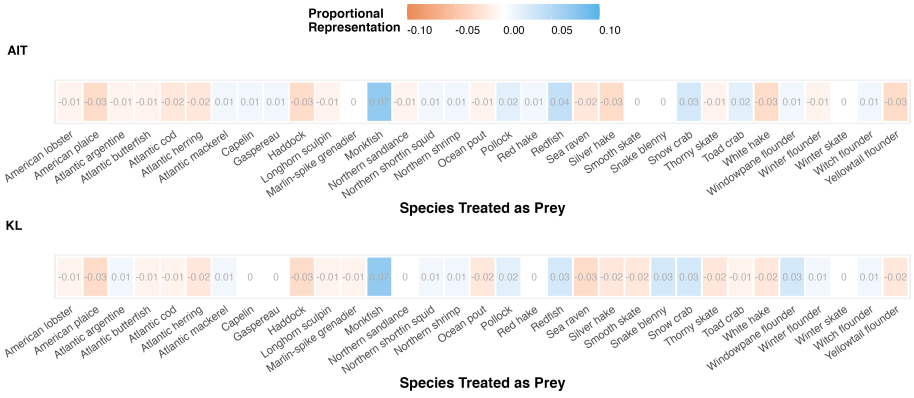


Figure 4.14: Proportional representation of each species in the extended prey library in prey-on-prey diet estimates above or below the expected proportion of $1/33 = 0.03$.

Chapter 5

Discussion and Conclusions

5.1 Discussion of Results

For the goal of correctly excluding prey, FS appears to be the best choice among the methods we evaluated as its match rates for true zero species were much higher across all diet groups; it is also the only appropriate choice when the number of potential prey species exceeds the number of dietary FAs. However, BE provides a conservative alternative for reducing the prey library if the number of species does not exceed the number of FAs. While BE offers an additional measure of protection against incorrectly excluding a species consumed by a single predator, which may be useful when working with more variable real-life data, it can be noted that in our simulation studies neither method excluded a true nonzero species in any of the samples.

In simulation studies, FS, with its tendency to select smaller models, performed best for those true diets with fewer nonzero species. Depending upon the predator, this may not be a significant limitation in practice; the number of prey species consumed by various groups of grey seals, for example, ranged from 2.4 to 5.5 in one analysis by Beck et al. (2007). That FS tended to select very few species

in the real-life examples without the implementation of a minimum model size is in line with those estimates. In these cases (with the number of species in the true diet ≤ 5), chi-squared distances between true and estimated diets were lower for FS estimates than other methods, particularly when the predators within a group were assigned distinct dietary proportions. This difference may be attributable to the fact that when many true zero species are correctly excluded in the selection process, the finite sum of each predator's diet composition is prevented from being divided amongst many nonzero species. By comparison, QFASA estimates in this setting were vulnerable to incorrectly assigning nonzero proportions to many species, contributing to greater distances between true and estimated diets.

Still, when additional prey were introduced bringing the number of species above the number of FAs, FS maintained its high classification accuracy in simulation studies. Across the four diets we evaluated, true nonzero species were classified correctly 100% of the time, and true zero species were correctly excluded an average of $85 - 95\%$ of the time. These results support the viability of FS as a means of selecting a prey library when the number of potential prey would otherwise be too great for valid estimation since there does not currently exist a FA-based method for reliably estimating the diets of predators in this case.

Some species appear to be more difficult to classify, but the difficulty does not appear to be directly proportional to the estimation difficulty outlined in Figure 4.13. Moreover, the difficulty of classifying certain species does not seem to be consistent across methods or diet groups, and suggests that classification accuracy may be influenced by the combination of species in a prey library as well as the true diets of the predators being analyzed. Indeed, we observed changes in the selected subsets of prey for the same samples of real-life seals when additional species were added to the prey library.

The high proportions of longhorn sculpin in FS diet estimates for the seals

sampled in 1997 do not seem to agree with previously published estimates for grey seal diets (Beck et al., 2007). While the estimates are similar to those obtained using QFASA [AIT], estimates for the same seals using QFASA [KL] favoured proportions of Atlantic cod and white hake at the exclusion of longhorn sculpin. Considering the equivalence laid out in section 3.1.3, it is not unusual for our QFASA MLE estimates to be more similar to QFASA [AIT] estimates. However, the analysis of our prey library in Figure 4.13 shows that QFASA [KL] may be more reliable for estimating Atlantic cod and, to a greater extent, white hake for this particular dataset.

For this predator sample, the pair of redfish and longhorn sculpin were selected as starting species as they produced the lowest AIC of all the possible starting pairs. It is noteworthy that the second best option was the pair of redfish and Atlantic cod with an AIC just 0.44% higher. The third best candidate was the pair of redfish and white hake with an AIC 5.32% higher. These AIC values suggest that multiple candidates fit similarly well, and further work may be needed to explore options for choosing among similar candidates by more sophisticated means than simply choosing the lowest AIC each time, particularly when the next best candidate is as close as in this example.

Addressing challenges with the selection of starting species more broadly is an important consideration for future work. Because the choice of starting species can affect the perceived contribution of the next species to the fit of the model, these runner-up species are often missed in later steps and ultimately excluded from the model. The current method of choosing the best pair for starting species presents another concern in that by limiting the model to two species, there is a risk that the best-fitting candidate may be one that includes species which approximate two or more different species in the true diet. Taking the same best subset approach while increasing the number of starting species may not be a viable solution as the number of candidate models to evaluate becomes much less feasible beyond two species. In

practice, these issues can be circumvented by specifying starting species based on prior knowledge.

5.2 Future Work

The FS algorithm could be further applied to allow for a greater number of within-species clusters to be treated as separate prey groups, which may be a worthwhile path to investigate as a means of improving estimation accuracy when prey data includes species whose mean FA signatures do not resemble the individuals.

Beyond the additional information criteria methods discussed above, the QFASA MLE method could be expanded for use in other variable selection techniques in this setting. Many techniques rely upon penalized likelihoods and could be adapted for this purpose. The QFASA MLE method could also be extended for use in estimating calibration coefficients instead of relying on captive feeding studies.

References

[1] C.A. Beck et al. "Sex differences in grey seal diet reflect seasonal variation in foraging behaviour and reproductive expenditure: evidence from quantitative fatty acid signature analysis". In: *Journal of Animal Ecology* 76 (2007), pp. 490–502.

[2] J.F. Bromaghin, S.M. Budge, and G.W. Thiemann. "Detect and exploit hidden structure in fatty acid signature data". In: *Ecosphere* 8 (2017), e01896.

[3] J.F. Bromaghin et al. "Assessing the robustness of quantitative fatty acid signature analysis to assumption violations". In: *Methods in Ecology and Evolution* 7 (2015), pp. 51–59.

[4] J.F. Bromaghin et al. "Simultaneous estimation of diet composition and calibration coefficients with fatty acid signature data". In: *Ecology and Evolution* 7 (2017). DOI: 10.1002/ece3.3179.

[5] H.W. Cook and C.R. McMaster. "Biochemistry of Lipids, Lipoproteins and Membranes". In: *Biochemistry of Lipids, Lipoproteins and Membranes*. Amsterdam, NL: Elsevier Science, 2002, pp. 181–202.

[6] J.J. Egozcue et al. "Isometric Logratio Transformations for Compositional Data Analysis". In: *Mathematical Geology* 35 (2003), pp. 279–300.

[7] C. Goetsch et al. "Energy-rich mesopelagic fishes revealed as a critical prey resource for a deep-diving predator using quanitative fatty acid signature analysis". In: *Frontiers in Marine Science* 5 (2018), p. 430.

[8] S.J. Iverson et al. "Quantitative fatty acid signature analysis: A new method of estimating predator diets". In: *Ecological Monographs* 74 (2004), pp. 211–235.

[9] J. M. McNichol. "Simultaneous Maximum Unified Fatty Acid Signature Analysis". MSc thesis. University of New Brunswick, 2022.

[10] D.L. Phillips and J.W. Gregg. "Source partitioning using stable isotopes: coping with too many sources". In: *Oecologia* 136 (2003), pp. 261–269.

[11] G. Schwarz. "Estimating the dimension of a model". In: *Annals of Statistics* 6 (1978), pp. 461–464.

[12] H. Steeves. "Maximum likelihood approach to diet estimation and inference based on fatty acid signatures". PhD thesis. Dalhousie University, 2020.

[13] C. Stewart. "An approach to measure distance between compositional diet estimates containing essential zeros". In: *Journal of Applied Statistics* 44 (2016), pp. 1137–1152.

[14] C. Stewart. *QFASA: Quantitative Fatty Acid Signature Analysis*. R package version 1.2.1. 2024. URL: `https://CRAN.R-project.org/package=QFASA`.

[15] C. Stewart and C. Field. "Managing the Essential Zeros in Quantitative Fatty Acid Signature Analysis". In: *Journal of Agricultural, Biological, and Environmental Statistics* 16 (2011), pp. 45–69.

[16] M. Tsagris and C. Stewart. "A Dirichlet Regression Model for Compositional Data with Zeros". In: *Lobachevskii Journal of Mathematics* 39 (2018), pp. 398–412.

Curriculum Vitae

Tyler Rideout

Education

University of New Brunswick Saint John, NB

MSc; Statistics Expected 2025

University of New Brunswick Saint John, NB

BSc; Double Major in Mathematics and Statistics 2020

Conference Presentations

Rideout, T., Stewart, C., and Stephenson, M. May 2022. **Prey selection for fatty acid signature analysis** [Conference presentation]. Canadian Statistics Student Conference.

Rideout, T., Stewart, C., and Stephenson, M. June 2024. **Prey selection for fatty acid signature analysis using the Akaike information criterion** [Conference presentation]. Statistical Society of Canada Annual Meeting (St. John's, NL).

Awards

Best Oral Presentations of the CSSC Award 2022