MAXIMUM LIKELIHOOD APPROACH TO DIET ESTIMATION AND INFERENCE BASED ON FATTY ACID SIGNATURES

by

Holly Steeves

Submitted in partial fulfillment of the requirements
for the degree of Doctor of Philosophy

at

Dalhousie University
Halifax, Nova Scotia
October 2020

Table of Contents

List of Tables

List of Figures

Abstract

Diet compositions of marine predators are often of interest for marine ecologists in trophic structure studies where non-lethal sampling has created a need for non-invasive diet estimation techniques. Methods using fatty acids have been developed to obtain dietary estimates that have previously been difficult to acquire. Building on the existing method, quantitative fatty acid signature analysis (QFASA), we have constructed a maximum likelihood approach to estimating dietary proportions.

This novel approach includes random effects to account for the unobserved prey that were consumed by the predator. Not only does it include variability of the prey and predator FA signatures in the model, but with the use of parametric bootstrapping, we can obtain confidence bounds on these diet estimates as well. These bounds will prove to be accurate for proportions away from the edges of the simplex.

It is also able to include covariates in the model. With use of a link function, the diet proportions are assumed to be a function of the covariates. The coefficients of this relationship are then optimized by using the same likelihood function as before, only subbing the link function in place of the diet proportions. This method yields a summary diet for all unique sets of covariates. It also allows for inference on diet estimates between various groups, such as sex, age or environmental factors. Simulations show that not only are the summary estimates accurate, but the inference leads to making the correct decision in all cases run.

Finally, these techniques are used to analyze two real life data sets. The first is a captive study of harbour seals, for which true diets are known. This shows us that our method is estimating as accurately as QFASA. The second is a study of grey seals off of Sable Island. For this set, sex and type of population growth on Sable are recorded for each seal, so the covariate method is applied here. In comparison to QFASA, our method appears to yield similar summary estimates, and the test yielded results in

agreement with the beliefs of biologists.

Acknowledgements

I would like to give a heartfelt thank you to my three supervisors, Chris, Aaron and Connie, without whom I never would have made it this far. There were tears, blunders, curse words, but also laughter and breakthroughs. The knowledge I have learned from you will stay with me throughout my whole career, and I cannot thank you enough for that. I would also like to thank Stu Carson, for all the lunchtime beers that got me through the most stressful times in this process. Having some drinks and venting about it all was a true lifesaver. And of course, I need to thank my mom, whose nagging through my grade school years instilled in me my study habits and work ethic that got me here (no matter how much I hated it at the time).

There are so many others that I need to thank. Robyn, for always being there for me in any way I need. Adam and Heather, for all those visits that got me out of my head for a bit, and for producing the best nieces a girl could ask for. Jess and Kyle, for board game nights and cat sitting. Celina, for surviving quarantine with me. Archimedes and Fisher, for keeping me calm and distracting me in all the best ways.

Finally, I would like to thank NSERC, for supporting me through this journey financially for the first several years. Also, thank you to the Lett family, for supporting me in the final years when all the funding ran out. Without these funds, I never would have been able to devote my time and energy to this thesis.

Chapter 1

Introduction

Fisheries play an important role in the diets and livelihoods of people around the globe. Approximately 1.5 billion people rely on fish for at least 20% of protein intake, and 520 million people are supported by this trade (Badjeck et al. (2010)). Mismanaged fisheries have been estimated to cause a loss of economic benefits worth $50 million US annually (Kelleher et al. (2009)). It is estimated that 70% of global fisheries stocks are overexploited or have collapsed (Pauly et al. (2008)), which often leads to a decline in marine species abundance (Froese & Kesner-Reyes (2002)). Therefore, sustainability of these fisheries is an important issue, not only for humans, but also for marine ecosystems.

Attaining sustainable fisheries is becoming increasingly difficult due to the looming threat of climate change (MacNeil et al. (2010)). Climate change is causing loss of habitat and moving species to areas newly within their thermal tolerance range (Cheung et al. (2010)). Further understanding of the effects of climate change is required so that fisheries management can adapt to changes in distribution, productivity and resilience in fish stocks brought on by climate change (Brander (2010)). Marine ecosystems at the equator and the poles are being disproportionally affected by these changes due to the truncation and expansion of thermal limits (MacNeil et al. (2010)). Such major changes in ecosystem function and the consequent availability of prey will disrupt food webs to an unknown degree, with potentially important consequences for upper trophic level species responding to a changing prey base. Therefore, valuable information on the effects of ocean warming in marine food webs might be revealed through the study of predators' dietary habits.

Food webs provide a framework that links population dynamics, community structure and ecosystem processes together (Kaunzinger & Morin (1998)). Therefore, an

understanding of the trophic system, that is, the flow of energy between organisms in an ecosystem through consumption, can yield important information on the marine ecosystem as a whole. Through the study of food webs, an understanding of species' roles within an ecosystem can be obtained (Thapanand et al. (2009)). Food webs are known to respond to species invasions (Vander Zanden et al. (1999)), environmental disturbances (Wootton et al. (1996)) and global warming (Petchey et al. (1999)). Therefore, trophic dynamics and feeding relationships between species may provide important insights into how species are adapting to large scale ecosystem changes.

While feeding habits and diets for many species can be estimated from direct observation, most marine predators feed below the surface and therefore cannot be easily observed. In such cases, stomach contents have previously been used to identify diet compositions of predators (Hyslop (1980)) but this approach is not favoured since animals often need to be euthanized for analysis (Beckmann et al. (2013)). Euthanasia is not ideal, especially for endangered or protected species. Also, stomach contents do not identify a long term trend in diet, but are only suggestive of the most recent meal. Lastly, because soft tissues easily degrade in the stomach, a bias exists towards digestion-resistant hard parts. Therefore some soft-bodied prey species in the diet may not be found in stomach contents (Iverson et al. (2004a)). These problems create the desire for a method that is less intrusive, recovers longer term trends and is capable of identifying both hard and soft bodied prey species.

Problems associated with stomach content analysis have been widely recognized and as a result, there has been considerable interest in bulk stable carbon and nitrogen isotopes for characterizing short and long term trophic relationships (Gannes et al. (1997)). Stable carbon and nitrogen isotope analysis is a non-invasive procedure that can provide useful information about a predator's placement in the food web. Stable nitrogen isotope abundances in liver, muscle tissue and bone collagen yield information about the trophic level while stable carbon isotope abundances in muscle tissue give information about the primary production source (Hobson (1993)). Although bulk isotopes provide valuable trophic information, they are not typically useful for estimating the composition of consumer diets (Hobson (1993)).

1.1 Fatty Acid Signature Analysis

In order to estimate diet compositions of predators, a new method using fatty acids (FAs) was developed. Fatty acids are fundamental components of lipids or fat that are used to store energy and do not degrade during digestion (Iverson et al. (2004a)). Some FAs are absorbed into fat stores, or adipose tissues with little modification from their composition in consumed prey. Therefore FAs can be divided into dietary and non-dietary components based on whether or not they are biosynthesized within a predator. Those that are strictly obtained from diet are stored in the predator's tissue with hardly any metabolization and are therefore in very similar proportions to those in the consumed prey. As such, they can be used to estimate the diet composition of the predator using quantitative fatty acid signature analysis (QFASA; Iverson et al. (2004a)) by comparing the FA signature of the predator to the FA signatures of various potential prey.

The original QFASA method implements a distance (between compositions) minimization algorithm, the details of which will be discussed in Section 3.1. First, the diet of the predator is writen as a linear combination of the mean FA signatures of the prey. Then, in order to estimate diet proportions, the distance between the observed FA signature of the predator and the linear combination is minimized. Because of the compositional nature of both FA signatures and the diet vector (the elements represent proportions of a whole), standard multivariate statistical analysis, including Euclidean distance, are unsuitable. Originally, transformations based on log-ratios were proposed to bring the compositions into real space, however, FA data often includes zeros, making both ratios and logarithms impractical. In addition, FA data often has larger dimension than sample sizes, creating issues for estimation of parameters, including but not limited to identifiability problems. The simple nature of QFASA's model allows for adaptations, including which distance measure to use, which allows new measures to be included, which may better accommodate the restrictive compositions. Aitchison's distance is the recommended approach to measure distance between the compositional vectors in QFASA as it yields diet estimates

with the least bias, and smallest root mean squared error (Bromaghin et al. (2015)), however, in the presence of zeros, the chi-squared distance measure (Stewart & Field (2011)) may be a better choice.

The "*p.QFASA*" function in the QFASA R package (Iverson et al. (2004b)) computes the QFASA diet estimates for a given choice of distance. While QFASA is a useful method for estimating diet proportions, some biological concerns have led to improvements on the methodology. First, it is known that certain FAs are metabolized within the predator, and therefore the proportions will not exactly match those of the consumed prey. This means proportions of certain FAs in the predator will always be higher or lower than the proportions in the prey species (Kirsch et al. (2000)). To account for this, feeding experiments can be performed to obtain calibration coefficients which quantitatively account for this metabolization. Another biological improvement on QFASA is found by incorporating fat content into the estimation. Fattier prey species will contribute more to FA signatures of predators than less fatty species, therefore taking fat content into account improves estimation of diet proportions. The details of calibration coefficients and fat content are outlined in Section 3.1.

In addition to calibration coefficients and fat content, using specific FA subsets can help improve diet estimation. Not all FAs are included in QFASA, but typically two different subsets are considered, which vary from predator to predator. These are referred to as the extended dietary subset and the dietary subset. The extended dietary subset contains FAs that are influenced by both diet and biosynthesis, and the dietary subset contain FAs that are influenced by diet only. Since the dietary subset is not biosynthesized, the proportions of FAs will be absorted with little to no modification into the predator, and therefore we will obtain more accurate diet estimates. Originally, these subsets were extracted by selecting the FAs to be included from the FA signature, and rescale the signature to sum to 1. However, Bromaghin et al. (2016) argues rescaling the signatures distorts predator-prey relationships and could lead to a bias in diet estimation. He proposes an augmented matrix method of dealing with the subsets where the signatures are not rescaled, but an element

containing the remaining proportion of 1 minus the partial sum of FAs is added to the signature. This new method was found to be important if the partial sums significantly differ between prey types.

Although QFASA is widely used and accepted, it has some drawbacks. Firstly, it does not easily allow for inference, specifically confidence bounds or hypothesis testing (Stewart & Field (2011), Stewart et al. (2014)). Although there have been several different techniques proposed for testing for differences among diets, these require either obtaining diet estimates before the test (Stewart et al. (2014)) or are performed on fatty acids, and differences in diets are inferred (Steeves et al. (2016)). Until now, there has not been a way to integrate inference into the estimation process of QFASA, allowing for testing, standard errors, and confidence intervals from the estimation model. In this thesis, we propose a novel way of adding covariates into the estimation process, allowing for many improvements over the standard inference techniques for QFASA. Our novel method not only provides simultaneous estimation of diet proportions and modelling of covariates, but also uses parametric bootstrapping to estimate the standard error of the diet estimates and covariate coefficients. Using the bootstrapped estimates, we are able to test for differences in diets between groups, such as age, location, time periods, and sex, in the estimation of diet composition. This allows us to obtain more accurate summary diets among groups, such as pups, adolescents and adults, or males and females, as it is theorized that these groups will not be consuming the same diets. In addition, quantifying changes in predator diet among these groups, specifically through space and time, can help biologists understand changes to a prey base responding to warming oceans attributed to climate change.

Another issue that QFASA does not address is that prey that are consumed by the predator are not the same prey sampled in the preybase. Therefore, the sample of prey may not be a good representation of the consumed prey. The maximum likelihood (ML) procedure proposed here addresses the problem with the use of random effects. The true consumed prey FA signatures are modelled as unobserved random effects and the sampled prey help provide information about the distribution of these

unknown FA signatures. Although this adds complexity to the model, it allows us to more accurately depict what is happening in nature compared to QFASA.

With FA data, comes high dimensionality, but relatively small sample sizes. Often, 67 FAs are measured for 4-24 prey species, but less than 30 predator FA signatures are sampled. This poses a concern for usual multivariate analyses and estimation as standard multivariate models on small data sets with high dimensionality can lead to unstable coefficient estimates, inflated standard errors, reduced power and inaccurate conclusions (Bühlmann & Van De Geer (2011), Finch & Finch (2017)). FA data also include zeros, which make the usual logarithmic transformations (described in Section 2.1) an issue. Our novel ML method uses an updated form of approximations from Aitchison & Bacon-Shone (1999) for such complicated data sets incorporating the most recent and recommended transformation, the isometric log-ratio (ilr) transformation (see Definition 2.11). This is valuable not only to our methodology here, but any other statistical methods for compositional data that require a convex linear combination and the ilr transformation.

Within this thesis, the new ML methodology for diet estimation, both with and without covariates, will be explored via simulations and real life data. For both simulations and real life data, a prey base is required which includes FA signatures from a sample of plausible prey species. This prey base must be applicable to the predator species, location, and season that is being analysed. Therefore for this work, we use 3 different preybases: the spring Scotia shelf preybase for simulations, as it has the largest sample sizes, the Vancouver Aquarium preybase for the first real life study, as it includes samples from the containers of fish that were fed to the captive animals, and the winter Scotian shelf preybase for the grey seal set, as it is sampled in the same area and season as the seals.

Prey Base

The first preybase was collected off the Scotian shelf in spring, summer or fall of 1993-1996, and 1999 and is comprised of 1689 FA signatures using 67 FAs from 21 different prey species, as well as fat content measured as percent wet mass for each

individual. The second preybase was collected at the Vancouver Aquarium in 2003 as part of a captive study. This preybase has 307 FA signatures comprised of 67 FA, with 11 prey species, divided into 12 distinct prey groups, and a fat content measurement for each prey individual, measured as percent wet mass. The third prey base was collected off the Scotian shelf in spring, summer, or fall between 1990 and 2001, or in the Gulf of St. Lawrence from research cruises and commercial fisheries between 2002 and 2004. It is comprised of 1735 FA signatures using 67 FAs from 21 different prey species, and fat content measured as percent wet mass for each individual. Some prey species (American plaice, Atlantic butterfish, Atlantic herring, capelin and longhorn sculpin) were subdivided into smaller clusters, some based on size, others on seasonal variation. The collection details of these sets are described in Sections 4.1, 6.1 and 6.3.1, respectively.

Simulations

Simulations are necessary to ensure that the statistical methods are behaving properly. With simulations, all parameters are set, and data is generated based on those parameters. Then, the methods are used to estimate the "unknown" parameters, and estimates and true values can be compared. For our methods, we will call the generated data "pseudo-predators".

To simulate non-parametric pseudo-predators, a bootstrap sample of prey FA signatures is first collected from each prey species. The mean (or median) FA signatures of bootstrapped samples are calculated for each prey species. A linear combination of the mean summary prey signatures is then taken with the true diet. Adding error to the diets requires care and we carry it out by adding errors in the transformed scale. The error is back transformed and perturbed with the linear combination and the signature that results is our pseudo-predator FA signature. Alternatively, parametric pseudo-predators can be created by randomly generating a transformed prey FA signature for each prey species sampled from the multivariate normal distribution and then back transforming. This yields the summary FA signatures for the prey, and then the remaining steps are the same as above. This process is explained in more

detail in Sections 4.2 and 4.2.

True diets were selected using the function "make_diet_grid" in the package "qfasar" (Bromaghin (2017)). This function creates a grid of compositions that are equally spaced within the simplex. This allows us to explore the behaviours of methods over the entire space. See Table 4.4 for the diets that we used.

Experimental Studies

In addition to simulations, our methods are run on experimental data to examine the behaviour of the estimates on real life data. Two real life datasets are used for our study: one without covariates, but with known "true" diets (Vancouver Aquarium data set), and one with covariates but unknown diet (winter grey seal dataset).

The Vancouver Aquarium dataset used is from a captive feeding study on newly weaned harbour seals, collected by Nordstrom et al. (2008). Seals were split into three groups and fed either Pacific herring (*Clupea pallasii*) for 42 days, surf smelt (*Hypomesus pretiosus*) for 42 days, or fed Pacific herring for 21 days followed by surf smelt for 21 days. Blubber samples were collected at day 0, 21, and 42 in order to obtain FA signatures for analysis.

The winter grey seal dataset contains FA signatures for 502 adult grey seal blubber samples collected in winter between 1994 and 2015. These seals were sampled during the annual breeding season (December-January) on Sable Island, NS, and contained 183 males and 319 females. Also included in the data set is age at the time of sampling (if known), year group (3) referring to which period of population growth is occuring on Sable Island, and cohort which is the year the seal was born.

1.2 Thesis Overview

Though QFASA has been widely used and accepted, the lack of inclusion of variability of FAs in both predator and prey in the model has been a major limitation. The

original QFASA model only relies on the mean FA signatures of observed prey species so an important, yet non-trivial question exists: how do we incorporate the variability of prey FAs into the model, as well as recognize that the observed prey FAs are not the same prey FAs that are consumed by the predators? This thesis will introduce several novel models for doing just that, which rely on maximum likelihood estimation to obtain diet estimates for the predators. This methodology allows for other important improvements over QFASA, such as parametric bootstrapping to obtain standard errors of the estimates, robustness, and the option to include covariates, such as sex, age, season, or year, into the model to improve diet estimation and to test for the effects of covariates on diet.

Chapter 2 presents an introduction to compositional data analysis. This chapter includes the usual ways of dealing with compositions for analysis, such as logarithm based transformations and will discuss the pros and cons of such techniques. Methods for handling zeros, and parametric modelling will also be discussed, and challenges when working with compositional data will become apparent. One challenge that is relevant to the models that will be proposed, is determining the distribution of a convex linear transformation of compositions. Several approximations proposed in Aitchison & Bacon-Shone (1999) are explained and further modifications to suit our data are discussed. As we will be introducing novel ways to model and perform inference on compositional data, several techniques for inference which have been used previously are presented in this chapter as well.

Chapter 3 describes the existing method, QFASA, in detail, and it's limitations are made apparent. The solution in the form of a maximum likelihood approach is proposed to improve upon the methodology of QFASA. This new model takes a linear transformation of the unknown diet proportions for each prey species, with a random effect, representing the unobserved FA signatures of prey that are consumed by the predator. The theory for this technique is introduced, and the implementation of the model is detailed.

Chapter 4 presents simulations to assess how well our MLE method preforms, particularly in comparison to QFASA, for the diet estimates. Simulation is done by creating pseudo-predators, either parametrically using a multivariate normal approximation to the transformed prey base or non-parametrically by sampling from the prey base, and a "true" diet. Twenty diets spanning the simplex, along with three different groups of 4 species were used for assessment. Results are given for the diet estimates along with their standard errors.

Chapter 5 discussed the important contribution of the use of covariates in the model for more accurate estimation of diet composition and the testing of the effect of covariates on diet. The theory and methodology is described here, as well as the simulations performed. Two groups (male/female) of pseudo-predators are generated parametrically based on two distinct diets, and one summary diet is estimated for each group. This is run using the prey group with 4 species having highly different FA signatures, for all combinations of 10 distinct diets. The settings for the simulations are specified, and the results summarized.

Chapter 6 describes the data sets and feeding experiments used in the real life analysis. Analysing real life data on which the true information is known (such as diet proportions) is necessary to ensure methods are behaving as expected on true biological data. The results from these experimental studies are summarized and used to select the best model choices. In addition, a data set with unknown parameters is used to determine diet estimates across male/female groups, a variety of ages and 3 different periods of population growth of grey seals. This set is used to emphasize the relevance of the new ML methodology for inclusion of covariates and ability for parametric inference.

Finally, Chapter 7 contains the conclusions of this thesis. A summary of results and recommendations are presented, and future work in this area is also addressed.

Chapter 2

Compositional Data

Compositional data commonly arise in many disciplines, including geology, ecology, and chemistry. A composition depicts relative information through quantitative descriptions of parts of a whole. This imposes constraints on the data; specifically, that the elements are non-negative and sum to 1. These constraints create several difficulties when statistical analysis is needed. Aitchison (1986) warns of applying standard statistical techniques on compositional data without taking these constraints into account, stating it is improper and inadequate for the data, and leads to "dubious" conclusions such as misinterpretation of spurious correlations (Pearson (1897)). As a pioneer in the field of compositional data analysis, Aitchison presented several techniques for analysing this type of data, and many breakthroughs have been achieved since. In this section, I will present several definitions, notations, and methods pertaining to compositional data.

2.1 Definitions

Definition 2.1. *Aitchison (1986) defines a composition to be a vector* $\boldsymbol{u}_o = (u_{o1}, u_{o2}, ..., u_{oD})$ *that has non-negative elements and satisfies the unit-sum constraint. The unit-sum constraint refers to the elements of the vector* $\boldsymbol{u}_o$ *summing to one:*

$$u_{o1} + u_{o2} + ... + u_{oD} = 1$$

Definition 2.2. *The space on which a composition is defined is called the simplex,* $\mathcal{S}^d$*, where* $d = D - 1$*. It is defined by:*

$$\mathcal{S}^d = \{(u_{o1}, u_{o2}, ..., u_{oD}) | u_{o1} \geq 0, u_{o2} \geq 0, ..., u_{oD} \geq 0, u_{o1} + u_{o2} + ... + u_{oD} = 1\}$$

Any vector with positive (or non-negative) elements can be transformed to be a composition defined on the simplex, $\mathcal{S}^d$. This process can be explained through

Definitions 2.3 to 2.4.

Definition 2.3. *A basis $\boldsymbol{x}$ of D parts is a vector of length D with positive elements recorded on the same scale.*

Definition 2.4. *The closure operator, $\mathcal{C}$, transforms each vector $\boldsymbol{x}$ of positive elements onto the simplex, $\mathcal{S}^d$.*

$$\mathcal{C}(\boldsymbol{x}) = \frac{1}{\sum_{i=1}^{D} x_i}\boldsymbol{x}$$

Every basis $\mathbf{x}$ yields a unique composition $\mathbf{u}_o = \mathcal{C}(\mathbf{x})$, however the converse is not true. This means that there are many bases that correspond to the composition $\mathbf{u}_o$. Any positive scalar multiple of the composition $\mathbf{u}_o$, $\{a\mathbf{u}_o, a > 0\}$ is in fact a basis for $\mathbf{u}_o$.

Definition 2.5. *If S is any subset of the parts $u_{o1}, u_{o2}, ..., u_{oD}$ of a D-dimensional composition $\boldsymbol{u}_o$, and $\boldsymbol{u}_{oS}$ is the vector formed from the subset of parts S, then $\mathcal{C}(\boldsymbol{u}_{oS})$ is a subcomposition of $\boldsymbol{u}_o$.*

An important operation that exists in the d-dimensional simplex is called perturbation.

Definition 2.6. *Aitchison (1986) defines the operation $\boldsymbol{x}\circ$ as a perturbation which is a one-to-one transformation from $\mathcal{S}^d$ to $\mathcal{S}^d$ calculated by:*

$$\boldsymbol{x} \circ \boldsymbol{u}_o = \mathcal{C}(x_1 u_{o1}, ..., x_D u_{oD})$$

Here, $\boldsymbol{u}_o$ is a D-part composition being operated on by the perturbing vector $\boldsymbol{x}$, which is a vector of length D with positive elements to form the perturbed composition $\boldsymbol{x}\circ\boldsymbol{u}_o$. Also, $\mathcal{C}$ refers to the closure operator, defined in Definition 2.4 which is used to ensure that the vector will sum to 1.

The inverse of a perturbation is $\boldsymbol{x}^{-1}\circ$, where

$$\boldsymbol{x}^{-1} = (1/x_1, 1/x_2, ..., 1/x_D)$$

In Aitchison (1986), several transformations are presented which transform from the interior of the d-dimensional simplex $\mathcal{S}^d$ to d-dimensional real space $\mathcal{R}^d$ or $\mathcal{R}^D$. The first of which is the multiplicative logistic transformation.

Definition 2.7. *The multiplicative logistic transformation (ml) is a one-to-one transformation that moves compositional data from the d-dimensional simplex $\mathcal{S}^d$ to d-dimensional real space $\mathcal{R}^d$ using the following:*

$$u_{mi} = \log\left(\frac{u_{oi}}{1 - u_{o1} - ... - u_{oi}}\right), i = 1, ..., d$$

The inverse of the ml transformation is given by:

$$u_{0i} = \frac{e^{u_{mi}}}{(1 + e^{u_{m1}}) \cdots (1 + e^{u_{mi}})}, i = 1, ..., d$$

The multiplicative logistic transformation is not commonly used (except for parametric modelling in Definition 2.22) as it depends on the ordering of the parts. This transformation is not the only option for transforming compositions; several other transformations are available, including the additive log-ratio transformation (alr), the centred log-ratio transformation (clr) and the isometric log-ratio transformation (ilr).

Definition 2.8. *The additive log-ratio transformation (alr) is a one-to-one transformation that moves compositional data from the d-dimensional simplex $\mathcal{S}^d$ to d-dimensional real space $\mathcal{R}^d$ using the following:*

$$\boldsymbol{u}_a = alr(\boldsymbol{u}_o) = \left[\log\frac{u_{o1}}{u_{oD}}, ..., \log\frac{u_{od}}{u_{oD}}\right]$$

where $d = D - 1$. The inverse of this transformation is:

$$\boldsymbol{u}_o = alr^{-1}(\boldsymbol{u}_a) = \mathcal{C}[exp(u_{a1}, u_{a2}, ..., u_{ad}, 0)]$$

The alr transformation has several disadvantages including that it is asymmetric in the parts of its composition, and that it is not isometric; Aitchison's distance (common distance measured used for compositional data, defined in Definition 2.15) and angles in the simplex are not preserved into Euclidean space (Egozcue et al. (2003)). The

clr transformation, defined below, has advantages over the alr transformation since it is symmetric in its components, as well as isometric.

Definition 2.9. *The clr transformation is a symmetric transformation from $\mathcal{S}^d$, the d-dimensional simplex, to $\mathcal{U}^D$, a $D-1$-dimensional hyperplane of real space $\mathcal{R}^D$, defined in Aitchison (1986) as:*

$$\boldsymbol{u}_c = clr(\boldsymbol{u}_o) = \left[\log \frac{u_{o1}}{\hat{g}(\boldsymbol{u}_o)}, ..., \log \frac{u_{oD}}{\hat{g}(\boldsymbol{u}_o)}\right]$$

where $\hat{g}(\boldsymbol{u}_o) = (u_{o1} \cdots u_{oD})^{1/D}$ represents the empirical geometric mean of the composition, and

$$\mathcal{U}^D = \{[x_1, ..., x_D] : x_1 + x_2 + ... + x_D = 0\}$$

The inverse of the clr transformation can be found by:

$$\boldsymbol{u}_o = clr^{-1}(\boldsymbol{u}_c) = \mathcal{C}[exp(\boldsymbol{u}_c)]$$

Although the clr transformation has symmetry and isometry, it has several issues, including subcompositional incoherence (defined below), a singular variance matrix (Filzmoser et al. (2009)), and being constrained to a subspace (Egozcue et al. (2003)). If subcompositional coherence is not satisfied, then the subcompositions (Definition 2.5) are not behaving as orthogonal projections. That is, at least one of the properties in Definition 2.10 is not satisfied.

Definition 2.10. *Subcompositional coherence holds when both features below are met.*

1. *The distance between two full compositions must be greater than or equal to the distance between any two subcompositions.*

2. *The distance between two compositions remains the same when the two compositions are scaled by a constant.*

The singular variance matrix causes problems for analysis after transformation, as it is not invertible. This particularly causes problems when assuming normal distributions on the transformations as the density involves the inverse of the variance

matrix. The ilr transformation defined below is preferred when dealing with compositional data as not only is it symmetric and isometric like the clr transformation, but also is subcompositionally coherent, has a non-singular variance matrix, and is unconstrained.

Definition 2.11. *From Egozcue et al. (2003), the ilr transformation maps from the d-dimensional simplex, $\mathcal{S}^d$, to d-dimensional real space, $\mathcal{R}^d$. We can calculate the ilr transformation, $\boldsymbol{u}$ from the clr transformation, $\boldsymbol{u}_c$, by:*

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

where $\boldsymbol{V}$ is a $D \times d$ orthonormal basis of the clr-plane and $\boldsymbol{V}^T \cdot \boldsymbol{V} = \boldsymbol{I}_d$, $\boldsymbol{V} \cdot \boldsymbol{V}^T = \boldsymbol{I}_D - \frac{1}{D}\boldsymbol{J}_D$, and $\boldsymbol{j}_D^T \cdot \boldsymbol{V} = \boldsymbol{0}_d^T$.

Here, $\boldsymbol{I}_D$ represents the $D \times D$ identity matrix, $\boldsymbol{J}_D$ is a $D \times D$ matrix of 1s and $\boldsymbol{j}_D$ is column vector of 1s of length D, and $\boldsymbol{0}_d$ is a column vector of 0s of length d.
The inverse of this transformation is given by:

$$\boldsymbol{u}_o = ilr^{-1}(\boldsymbol{u}) = \mathcal{C}[exp(\boldsymbol{u} \cdot \boldsymbol{V}^T)]$$

There are as many different ilr transformations as there are bases for the clr-plane, but here we chose to use a version of $\mathbf{V}$ defined in Egozcue et al. (2003) which is based on the Helmert matrix. This is also the version of $\mathbf{V}$ that the function ilrBase() uses in the package "compositions" in R (van den Boogaart et al. (2014)). It is given by:

$$\mathbf{V} = \begin{pmatrix} \frac{-1}{\sqrt{2}} & \frac{-1}{\sqrt{6}} & \cdots & \frac{-1}{\sqrt{(D-1)(D-2)}} & \frac{-1}{\sqrt{D(D-1)}} \\ \frac{1}{\sqrt{2}} & \frac{-1}{\sqrt{6}} & \cdots & \frac{-1}{\sqrt{(D-1)(D-2)}} & \frac{-1}{\sqrt{D(D-1)}} \\ 0 & \frac{2}{\sqrt{6}} & \cdots & \frac{-1}{\sqrt{(D-1)(D-2)}} & \frac{-1}{\sqrt{D(D-1)}} \\ \vdots & \vdots & \ddots & \vdots & \vdots \\ 0 & 0 & \cdots & \frac{-1}{\sqrt{(D-1)(D-2)}} & \frac{-1}{\sqrt{D(D-1)}} \\ 0 & 0 & \cdots & \frac{D-2}{\sqrt{(D-1)(D-2)}} & \frac{-1}{\sqrt{D(D-1)}} \\ 0 & 0 & \cdots & 0 & \frac{D-1}{\sqrt{D(D-1)}} \end{pmatrix} \tag{2.1}$$

2.2 Issues with Compositional Data

2.2.1 Analysis for Compositional Data

Aitchison (1986) discusses at length the difficulties and concerns associated with compositional data analysis. One of these difficulties lies in the frequent high dimensionality of compositional data. This makes graphical interpretations difficult, as the human eye cannot easily see in more than 3 dimensions. Previously, projections onto 2 or 3 dimensional space have been analysed, however this has been subject to criticism as a partial analysis which focuses only on some subcomposition formed from the composition. The unit-sum constraint also poses problems when performing a graphical analysis. The patterns perceived in the space of the simplex (for example, on a ternary diagram) will not necessarily coincide with similar interpretations in real space.

Another significant difficulty is what Aitchison (1986) calls an absence of an interpretable covariance structure. Standard covariance and correlation matrices (in his text, these are referred to as "crude" covariance and correlation structures) are a cause for concern for interpretation. For example, consider the interdependence of the components of the composition, $\mathbf{u}_o$, below.

$$\begin{aligned} cov(u_{o1}, u_{o1} + u_{o2} + ... + u_{oD}) &= cov(u_{o1}, 1) \\ &= 0 \end{aligned}$$

since by definition, $u_{o1} + u_{o2} + ... + u_{oD} = 1$. Therefore,

$$\begin{aligned} 0 &= cov(u_{o1}, u_{o1} + u_{o2} + ... + u_{oD}) \\ &= cov(u_{o1}, u_{o1}) + cov(u_{o1}, u_{o2}) + ... + cov(u_{o1}, u_{oD}) \\ &= var(u_{o1}) + cov(u_{o1}, u_{o2}) + ... + cov(u_{o1}, u_{oD}) \end{aligned}$$

From this, we get:

$$cov(u_{o1}, u_{o2}) + ... + cov(u_{o1}, u_{oD}) = -var(u_{o1})$$

Therefore, at least one of $cov(u_{o1}, u_{oi}), i = 1, ..., D$ has to be negative. This also applies to the covariance between any other u_{oi} and the other components. This causes problems with interpretations since the correlations are not free to take on any value between (-1,1). This is referred to as the "negative bias difficulty".

Another issue with the covariance of components lies with the lack of relationship between the covariance matrix of a subcomposition and that of the full composition. Similarly, there is often an absence of a relationship between the covariance or correlation of a basis $\mathbf{x}$ and that of it's composition $\mathbf{u}_o = \mathcal{C}(\mathbf{x})$.

Many of the techniques Aitchison (1986) proposes to solve the difficulties associated with compositional data involve transforming the data from the restrictive simplex onto unconstrained real space. This is done through transformations such as the alr, clr or ilr transformations defined in Definitions 2.8, 2.9 and 2.11 respectively. These transformations rely on ratios and logarithms, both of which cause problems for data with zero elements. Even the distance measures and measures of centre recommended for compositional data such as those defined in Definitions 2.15, 2.18 and 2.15 involve ratios or logarithms, making analysis of compositional data involving zeros particularly difficult. Thus, there is an entire field of research devoted to determining how best to handle zeros in compositional data sets.

2.2.2 Dealing with zeros

Many compositional data sets involve zero elements, or proportions of 0 in the vectors. For example, with FA data, we often have proportions of FAs that are below the detection limit of our instruments and are therefore recorded as zero, or with a diet composition, a predator completely avoids eating a certain species, say squid, thus

0% of its diet is squid. However, in order to calculate the alr, clr or ilr transformation, logarithms and ratios of the elements in the composition are performed. Logarithms and ratios are a common method for dealing with compositional data, whether it be in distance measures or transformations, making zeros a problem when analysing compositional data. There are three types of zeros as described in Pawlowsky-Glahn & Buccianti (2011): count zeros, essential zeros and rounded zeros. Count zeros are zeros present in discrete data, or count data, that represent a true absence of the element in the data. Essential zeros are similar in that there is a true absence of the element, however they occur in continuous data. Essential zeros are sometimes also referred to as absolute zeros or structural zeros. Lastly, rounded zeros occur when the value of the element fell below a threshold (often due to instrumentation) and was then rounded to zero. These zeros could be considered as missing values, since the true value of the element was not observed. Rounded zeros are very common in many areas of compositional data as all instrumentation has a detection limit, and are also the easiest to deal with. Therefore, rounded zeros have the most techniques available for handling them (Pawlowsky-Glahn & Buccianti (2011), Aitchison (1986), Martín-Fernández & Thió-Henestrosa (2006)).

Martín-Fernández & Thió-Henestrosa (2006) developed three different techniques for replacing the rounded zeros in a composition. In these methods, they treat zeros as missing values and perform an imputation technique for filling in the missing values. The three techniques are referred to as additive replacement, simple replacement, and multiplicative replacement and with the formulas shown below, they yield the replaced composition $\mathbf{r} = (r_1, ..., r_D)$. δ_j below is the imputation value for the j^{th} element in the composition, $\mathbf{u}_o$ is a D-dimensional composition with Z zeros, and c is the sum constraint of the composition, usually 1 or 100%.

Definition 2.12. *The additive replacement method is an imputation technique for rounded zeros that are considered missing values. The replaced composition* $\boldsymbol{r} = (r_1, ..., r_D)$ *for* $\boldsymbol{u}_o$ *is calculated by:*

$$r_j = \begin{cases} \frac{\delta_j(Z+1)(D-Z)}{D^2} & \text{if } u_{oj} = 0 \\ u_{oj} - \frac{Z+1}{D^2}\left(\sum_{k|u_{ok}=0} \delta_k\right) & \text{if } u_{oj} > 0 \end{cases}$$

Definition 2.13. *Similar to the additive replacement method, the simple replacement method is another imputation technique for rounded zeros, where the replaced composition is calculated by:*

$$r_j = \begin{cases} \frac{c}{c+\sum_{k|u_{ok}=0} \delta_k}\delta_j & \text{if } u_{oj} = 0 \\ \frac{c}{c+\sum_{k|u_{ok}=0} \delta_k}u_{oj} & \text{if } u_{oj} > 0 \end{cases}$$

Definition 2.14. *The final imputation technique for rounded zeros is the multiplicative replacement method. It's replaced composition is calculated by:*

$$r_j = \begin{cases} \delta_j & \text{if } u_{oj} = 0 \\ \left(1 - \frac{\sum_{k|u_{ok}=0} \delta_k}{c}\right) u_{oj} & \text{if } u_{oj} > 0 \end{cases}$$

The user can decide upon an imputation value, however in Pawlowsky-Glahn & Buccianti (2011), they suggest using 65% of the rounding threshold as the imputation value. For example, if the j^{th} part can be measured to 0.2 units, the rounding threshold for x_j would be 0.1, and the suggested imputation value δ_j would be equal to 0.065. There have been some more recent developments on replacement methods such as the parametric technique proposed in Palarea-Albaladejo & Martín-Fernández (2008) that utilizes the EM algorithm to replace the rounded zeros in the composition. Another recent technique using multiplicative modification and log-normal probabilities is described in Palarea-Albaladejo & Martin-Fernandez (2013). Here, we have chosen to use the non-parametric multiplicative replacement method in our analyses as it is simple and fast to implement, and yields relatively good results (Martın-Fernandez et al. (2011).

2.3 Measures of Distance

Measuring distance between compositions in a logical and accurate way is an important part in many applications of compositional data. Techniques such as quantitative

fatty acid signature analysis (Section 3.1), permutation tests, clustering methods, and nonparametric multivariate analysis of variance (Steeves et al. (2016)), require a valid way to measure distance between two compositions. Since compositions are defined on the simplex, distance measures defined in Euclidean space should be used. Many different distance measures are defined on the simplex, so Aitchison (1992) presented 7 criteria that a distance measure should follow in order to be satisfactory. These criteria are as follows:

1. Positivity
   $f(\mathbf{u}_{o1}, \mathbf{u}_{o2}) > 0$ if $\mathbf{u}_{o1}$ and $\mathbf{u}_{o2}$ are not equivalent.

2. Zero difference between equivalent compositions.
   $f(\mathbf{u}_{o1}, \mathbf{u}_{o2}) = 0$ if $\mathbf{u}_{o1}$ and $\mathbf{u}_{o2}$ are equivalent ($u_{o1i} = u_{o2i}$ for all $i = 1, ..., D$).

3. Interchangeability of compositions.
   $f(\mathbf{u}_{o1}, \mathbf{u}_{o2}) = f(\mathbf{u}_{o2}, \mathbf{u}_{o1})$.

4. Scale invariance.
   $f(a\mathbf{u}_{o1}, A\mathbf{u}_{o2}) = f(\mathbf{u}_{o1}, \mathbf{u}_{o2})$ for every $a > 0$, $A > 0$.

5. Perturbation invariance.
   $f(q \circ \mathbf{u}_o, q \circ \mathbf{u}_{o2}) = f(\mathbf{u}_{o1}, \mathbf{u}_{o2})$ for every perturbation q.

6. Permutation invariance.
   $f(P\mathbf{u}_{o1}, P\mathbf{u}_{o2}) = f(\mathbf{u}_{o1}, \mathbf{u}_{o2})$ for every permutation P.

7. Subcompositional dominance.
   $f_D(\mathbf{u}_{o1}, \mathbf{u}_{o2}) \geq f_{D*}(\mathbf{u}^*_{o1}, \mathbf{u}^*_{o2})$, where $f_{D*}(\mathbf{u}^*_{o1}, \mathbf{u}^*_{o1})$ represents the distance between $D*$-dimensional subcompositions of $\mathbf{u}_{o1}$ and $\mathbf{u}_{o2}$, where $D* \leq D$.

The first three are necessary qualities for all distance measures and will hold for all defined below. The last 4 do not hold for all distance measures defined on the simplex, so we will consider only these when comparing the distance measures defined below.

The first measure we will consider is Aitchison's distance measure.

Definition 2.15. *Aitchison's distance between two compositions* $\boldsymbol{u}_{o1}$ *and* $\boldsymbol{u}_{o2}$*, of length* D*, defined in Martín-Fernández et al. (1998) is calculated as:*

$$AIT(\boldsymbol{u}_{o1}, \boldsymbol{u}_{o2}) = \left(\sum_{j=1}^{D} \left\{ \log \left[u_{o1j} / \hat{g}(\boldsymbol{u}_{o1}) \right] - \log \left[u_{o2j} / \hat{g}(\boldsymbol{u}_{o2}) \right] \right\}^2 \right)^{1/2}$$

where $\hat{g}$ refers to the empirical geometric mean defined as $\hat{g} = (\prod_{k=1}^{D} \mathrm{u}_{ok})^{1/D}$. Aitchison's distance satisfies all the criteria described above, however it is not suitable for compositions involving essential zeros. Angular and (crude) Mahalanobis distance measures are also defined in Martín-Fernández et al. (1998). Both of these measures satisfy all of Aitchison's criteria except for subcompositional dominance, however they are capable of handling zeros. Subcompositional dominance logically means that the addition of more information about the parts should never make the distance between two compositions smaller. This sounds desirable, however Stewart (2017) argues that for the application on compositional data with zeros, it is not a property that is of practical importance. Therefore, many of the distance measures we look at will ignore subcompositional dominance. The two measures are defined below.

Definition 2.16. *The Angular distance measure is defined as:*

$$ANG(\boldsymbol{u}_{o1}, \boldsymbol{u}_{o2}) = arccos \left(\sum_{j=1}^{D} \sqrt{\frac{u_{o1j}^2}{\sum u_{o1j}^2}} \sqrt{\frac{u_{o2j}^2}{\sum u_{o2j}^2}} \right)$$

Definition 2.17. *The (crude) Mahalanobis distance measure is defined as:*

$$MAH(\boldsymbol{u}_{o1}, \boldsymbol{u}_{o2}) = \left[(\boldsymbol{u}_{o1} - \boldsymbol{u}_{o2})^T \boldsymbol{K}^{+} (\boldsymbol{u}_{o1} - \boldsymbol{u}_{o2}) \right]^{1/2}$$

where $\mathbf{K}^{+}$ denotes the Moore-Penrose pseudo-inverse of the covariance matrix $\mathbf{K}$ for a compositional data set.

Another measure commonly used on compositional data, is the Kulback-Leibler distance. This distance measure is neither scale invariant, nor subcompositionally dominant, however it has proved useful in specific applications such as QFASA.

Definition 2.18. *The KL distance measure is defined as:*

$$KL(\boldsymbol{u}_{o1}, \boldsymbol{u}_{o2}) = \sum_{j=1}^{p} (u_{o1j} - u_{o2j}) \log \left(\frac{u_{o1j}}{u_{o2j}} \right)$$

Motivated by the desire for a statistical distance that satisfies subcompositional coherence and allows for handling zeros without modification or replacement, Stewart et al. (2014) proposed a chi-squared distance measure based on the work of Greenacre (2011), which is further explored in Stewart (2017). Chi-squared distance (Definition 2.19) maintains scale and permutation invariance, and does not require zeros to be changed, however it is not subcompositionally dominant. Once again, this is a property that we are not concerned about in this context.

Definition 2.19. *The chi-squared distance measure is defined as follows:*

$$CS(\boldsymbol{u_{o1}}, \boldsymbol{u_{o2}}) = \sqrt{2D} \left(\sum_{j=1}^{D} r_j \right)^{1/2} \tag{2.2}$$

where

$$r_j = \begin{cases} 0 & \text{if } u_{o1j} = u_{o2j} = 0 \\ \dfrac{\left(\frac{u_{o1j}}{\sum_{k=1}^{D} u_{o1k}} - \frac{u_{o2j}}{\sum_{k=1}^{D} u_{o2k}} \right)^2}{\frac{u_{o1j}}{\sum_{k=1}^{D} u_{o1k}} + \frac{u_{o2j}}{\sum_{k=1}^{D} u_{o2k}}} & \text{otherwise} \end{cases}$$

It should be noted that Stewart et al. (2014) proposes a more general definition of the chi-squared distance measure that relies on a power transformation parameter γ. The parameter was selected by considering decreasing values of γ until subcompositionally coherence was achieved, or nearly achieved. When Stewart (2017) explored this further, they found that the CS distance becomes unstable when there are many zeros present in the compositions, and that in this case, subcompositional incoherence was at a minimum when $\gamma = 1$. Thus, the more specific definition of the chi-squared distance shown above is used, where the γ parameter was dropped by setting $\gamma = 1$.

2.4 Parametric Models

Parametric modelling is another concern for compositions. The Dirichlet distribution, described below, is defined on $\mathcal{S}^d$ and is commonly used for compositional data.

However, in Aitchison (1986), he argues that the Dirichlet distribution is inadequate due to the strength of the independence structure of compositions modelled by this distribution. Therefore, Aitchison (1986) presents several other parametric models, such as the additive logistic Normal distribution and the multiplicative logistic Normal distribution, both based on one-to-one transformations from $\mathcal{S}^d$ to $\mathcal{R}^d$, defined below.

Definition 2.20. *Suppose* $\boldsymbol{U}_o = (U_{o1}, ..., U_{oD})$ *is a* D*-part composition.* $\boldsymbol{U}_o$ *is said to have a Dirichlet distribution with parameter* $\boldsymbol{\alpha} \in \mathcal{R}_+^D$, $\mathcal{D}^d(\boldsymbol{\alpha})$, *defined on* $\mathcal{S}^d$, *if it's density function is as follows:*

$$\frac{\Gamma(\alpha_1 + ... + \alpha_D)}{\Gamma(\alpha_1) \cdots \Gamma(\alpha_D)} u_{o1}^{\alpha_1 - 1} \cdots u_{od}^{\alpha_d - 1} (1 - u_{o1} - \cdots - u_{od})^{\alpha_D - 1}$$

If $\mathbf{u}_{\mathrm{o}}$ follows a Dirchlet distribution with parameter $\boldsymbol{\alpha}$, then:

- $E(\mathbf{U}_{\mathrm{oi}}) = \alpha_i / \alpha_+$
- $var(\mathbf{U}_{\mathrm{oi}}) = \alpha_i(\alpha_+ - \alpha_i)/[\alpha_+^2(\alpha_+ + 1)]$
- $cov(\mathbf{U}_{\mathrm{oi}}, \mathbf{U}_{\mathrm{oj}}) = -\alpha_i \alpha_j / [\alpha_+^2(\alpha_+ + 1)], (i \neq j)$
- $corr(\mathbf{U}_{\mathrm{oi}}, \mathbf{U}_{\mathrm{oj}}) = -(\alpha_i \alpha_j)^{1/2}[(\alpha_+ - \alpha_i)(\alpha_+ - \alpha_j)]^{-1/2}, (i \neq j)$

where $\alpha_+ = \alpha_1 + ... + \alpha_D$.

Definition 2.21. *Suppose* $\boldsymbol{U}_o = (U_{o1}, ..., U_{oD})$ *is a* D*-part composition. The composition* $\boldsymbol{U}_o$ *is said to have an additive logistic Normal distribution,* $\mathcal{L}^d(\boldsymbol{\mu}_a, \boldsymbol{\Sigma}_a)$, *if the additive log-ratio transformation of* $\boldsymbol{U}_o$, $\boldsymbol{U}_a$, *follows a Normal distribution,* $MVN^d(\boldsymbol{\mu}_a, \boldsymbol{\Sigma}_a)$, *with mean* $\boldsymbol{\mu}_a$ *and covariance matrix* $\boldsymbol{\Sigma}_a$, *where* $d = D - 1$.

That is, if $\boldsymbol{U}_0$ *follows an additive logistic Normal distribution,* $\mathcal{L}^d(\boldsymbol{\mu}_a, \boldsymbol{\Sigma}_a)$, *then its density function is given by:*

$$f(\boldsymbol{u}_0) = \frac{1}{(2\pi)^{d/2} |\boldsymbol{\Sigma}_a|^{1/2} (u_{o1} \cdots u_{oD})} \times \exp\left[-\frac{1}{2} \left(\log\left(\frac{\boldsymbol{u}_{o-D}}{u_{oD}} \right) - \boldsymbol{\mu}_a \right)^T \boldsymbol{\Sigma}_a^{-1} \left(\log\left(\frac{\boldsymbol{u}_{o-D}}{u_{oD}} \right) - \boldsymbol{\mu}_a \right) \right]$$

where $(u_{o1} \cdots u_{oD})^{-1}$ *is the Jacobian of the transformation and* $\boldsymbol{u}_{o-D}$ *represents the* $\boldsymbol{u}_o$ *composition without the* D^{th} *entry,* $\boldsymbol{u}_{o-D} = (u_{o1}, u_{o2}, ..., u_{o(D-1)})$.

As described earlier, the alr transformation is not the only one-to-one transformation from $\mathcal{S}^d$ to $\mathcal{R}^d$. So, another normality class similar to the additive logistic normal distribution arises.

Definition 2.22. *Suppose* $\boldsymbol{U}_o = (U_{o1}, ..., U_{oD})$ *is a D-part composition. The composition* $\boldsymbol{U}_o$ *is said to have a multiplicative logistic normal distribution,* $\mathcal{M}^d(\boldsymbol{\mu}_m, \boldsymbol{\Sigma}_m)$, *if* $\boldsymbol{U}_m$ *, the multiplicative logistic transformation of* $\boldsymbol{U}_o$, *defined in Definition 2.7 has a* $MVN^d(\boldsymbol{\mu}_m, \boldsymbol{\Sigma}_m)$ *distribution, with* $d = D - 1$.

That is, if $\boldsymbol{U}_0$ *follows a multiplicative logistic normal distribution,* $\mathcal{M}^d(\boldsymbol{\mu}_m, \boldsymbol{\Sigma}_m)$, *if its density is given by:*

$$f(\boldsymbol{u}_0) = \frac{1}{(2\pi)^{d/2}|\boldsymbol{\Sigma}_m|^{1/2}(u_{o1} \cdots u_{oD})} \times$$
$$\exp\left[-\frac{1}{2}\left\{\left(\log\left[\frac{u_{o1}}{1-\sum_{i=1}^{1} u_{0i}}\right], ..., \log\left[\frac{u_{od}}{1-\sum_{i=1}^{d} u_{0i}}\right]\right) - \boldsymbol{\mu}_m\right\}^T \times \right.$$
$$\left. \boldsymbol{\Sigma}_m^{-1}\left\{\left(\log\left[\frac{u_{o1}}{1-\sum_{i=1}^{1} u_{0i}}\right], ..., \log\left[\frac{u_{od}}{1-\sum_{i=1}^{d} u_{0i}}\right]\right) - \boldsymbol{\mu}_m\right\}\right]$$

where $(u_{o1} \cdots u_{oD})^{-1}$ *is the Jacobian of the transformation.*

Aitchison (1986) outlines many properties that result from the connection between $\mathcal{L}^d$ and $\mathcal{M}^d$ with MVN^d. First, several properties hold for $\mathcal{L}^d$, including the permutation property, the perturbation property and the subcompositional property described below.

Property 2.1. *Suppose the D-part composition,* $\boldsymbol{U}_0$, *distributed as* $\mathcal{L}^d(\boldsymbol{\mu}_a, \boldsymbol{\Sigma}_a)$, *is perturbed by a vector* $\boldsymbol{x}$ *of D positive components and is independent of* $\boldsymbol{U}_0$. *Then the distribution of the perturbed vector* $\boldsymbol{X} = \boldsymbol{x} \circ \boldsymbol{U}_0$ *is given below:*

Distribution of $\boldsymbol{x}$	*Distribution of* $\boldsymbol{X}$
$\mathcal{L}^d(\boldsymbol{\theta}, \boldsymbol{\Theta})$	$\mathcal{L}^d(\boldsymbol{\mu}_a + \boldsymbol{\theta}, \boldsymbol{\Sigma}_a + \boldsymbol{\Theta})$
Constant vector	$\mathcal{L}^d(\boldsymbol{\mu}_a + alr(\boldsymbol{x}), \boldsymbol{\Sigma}_a)$

Property 2.2. *Suppose the D-part composition, $\boldsymbol{U}_0$, is distributed as $\mathcal{L}^d(\boldsymbol{\mu}_a, \boldsymbol{\Sigma}_a)$ and if $\boldsymbol{U}_p = \boldsymbol{P}\boldsymbol{U}_0$ is the composition with the parts reordered by the permutation matrix $\boldsymbol{P}$, then $\boldsymbol{U}_P$ is distributed as $\mathcal{L}^d(\boldsymbol{\mu}_P, \boldsymbol{\Sigma}_P)$ where*

$$\boldsymbol{\mu}_P = \boldsymbol{Q}_P \boldsymbol{\mu}_a$$

$$\boldsymbol{\Sigma}_P = \boldsymbol{Q}_P \boldsymbol{\Sigma}_a \boldsymbol{Q}_P^T$$

$$\boldsymbol{Q}_P = \boldsymbol{F}\boldsymbol{P}\boldsymbol{F}^T\boldsymbol{H}^{-1}$$

and where $\boldsymbol{F}_{d,D} = [\boldsymbol{I}_d : -\boldsymbol{j}_d]$, $\boldsymbol{H} = \boldsymbol{I}_d + \boldsymbol{J}_d$, $\boldsymbol{I}_d$ is the d-dimensional identity matrix, $\boldsymbol{j}_d$ is a column vector of units, and $\boldsymbol{J}_d$ is a d-dimensional matrix of units.

$\mathbf{Q}_p$ is so defined due to the relationships between the ilr and clr transformations. $\mathbf{U}_p = \mathbf{P}\mathbf{U}_o$ is a simple reordering of the parts. Since the alr transformation depends on which part is last, it is not so simple. However, the clr transformation does not depend on ordering at all, therefore we can similarly state that $\mathbf{U}_{cp} = \mathbf{P}\mathbf{U}_c$. It can be shown that in order to convert from the alr transformation to the clr transformation, we can use $\mathbf{U}_a = \mathbf{F}\mathbf{U}_c$. To transform in the opposite direction, we get $\mathbf{U}_c = \mathbf{F}^T\mathbf{H}^{-1}\mathbf{U}_a$. So, we can apply these relationships, along with the relationships described above, to obtain:

$$\begin{aligned}\mathbf{U}_{ap} &= \mathbf{F}\mathbf{U}_{cp} \\ &= \mathbf{F}\mathbf{P}\mathbf{U}_c \\ &= \mathbf{F}\mathbf{P}\mathbf{F}^T\mathbf{H}^{-1}\mathbf{U}_a\end{aligned}$$

Thus, the definition of $\mathbf{Q}_P$ becomes $\mathbf{Q}_P = \mathbf{F}\mathbf{P}\mathbf{F}^T\mathbf{H}^{-1}$.

Property 2.3. *Suppose the D-part composition, $\boldsymbol{U}_0$, is distributed as $\mathcal{L}^d(\boldsymbol{\mu}_a, \boldsymbol{\Sigma}_a)$, and $\boldsymbol{U}_s = \mathcal{C}(\boldsymbol{S}\boldsymbol{x})$ is the subcomposition with parts selected with the $C \times D$ selecting matrix $\boldsymbol{S}$, then $\boldsymbol{U}_s$ is distributed as $\mathcal{L}^c(\boldsymbol{\mu}_s, \boldsymbol{\Sigma}_s)$, where*

$$\boldsymbol{\mu}_s = \boldsymbol{Q}_s \boldsymbol{\mu}_a$$

$$\boldsymbol{\Sigma}_s = \boldsymbol{Q}_s \boldsymbol{\Sigma}_a \boldsymbol{Q}_s^T$$

$$\boldsymbol{Q_s} = \boldsymbol{F}_{c,C}\boldsymbol{S}\boldsymbol{F}_{d,D}^{T}\boldsymbol{H}^{-1}$$

and where $c = C - 1$.

Two properties pertaining to the $\mathcal{M}^d(\boldsymbol{\mu}_m, \boldsymbol{\Sigma}_m)$ distribution are described below. Because of these properties, and the fact that $\mathcal{M}^d(\boldsymbol{\mu}_m, \boldsymbol{\Sigma}_m)$ is clearly dependent on the ordering of parts, it has an advantage over the $\mathcal{L}^d(\boldsymbol{\mu}_{\mathrm{a}}, \boldsymbol{\Sigma}_{\mathrm{a}})$ distribution when it comes to analysis with ordered parts, as $\mathcal{L}^d(\boldsymbol{\mu}_{\mathrm{a}}, \boldsymbol{\Sigma}_{\mathrm{a}})$ only depends on the order of which element is the last element. Most notably the properties listed below:

Property 2.4. *Suppose the D-part composition* $\boldsymbol{U}_0$ *follows a* $\mathcal{M}^d(\boldsymbol{\mu}_m, \boldsymbol{\Sigma}_m)$ *distribution, and consider the (c, D-c) partition of* $\boldsymbol{\mu}_m$ *and* $\boldsymbol{\Sigma}_m$*:*

$$\begin{bmatrix} \boldsymbol{\mu}_{m1} \\ \boldsymbol{\mu}_{m2} \end{bmatrix}, \qquad \begin{bmatrix} \boldsymbol{\Sigma}_{m11} & \boldsymbol{\Sigma}_{m12} \\ \boldsymbol{\Sigma}_{m21} & \boldsymbol{\Sigma}_{m22} \end{bmatrix}$$

Then the following properties hold:

1. *The amalgamation of* $(\boldsymbol{U}_o^{(c)}, \boldsymbol{j}_{D-c}^{T}\boldsymbol{U}_{o(c)})$ *is distributed as* $\mathcal{M}^c(\boldsymbol{\mu}_{m1}, \boldsymbol{\Sigma}_{m11})$*, where* $\boldsymbol{j}_{D-c}^{T}$ *= [1,1,...,1] of length* $D-c$*,* $\boldsymbol{U}_o^{(c)}$ *is the full first c parts of the composition, and* $\boldsymbol{U}_{o(c)}$ *is the final* $D-c$ *parts of the composition.*

2. *The subcomposition* $\mathcal{C}(\boldsymbol{U}_{o(c)})$ *is distributed as* $\mathcal{M}^{D-c}(\boldsymbol{\mu}_{m2}, \boldsymbol{\Sigma}_{m22})$.

Although Aitchison (1986) warned that the multiplicative logistic distributions shown above do not have as many nice properties as the additive logistic distributions, Stewart & Field (2011) found the multiplicative logistic distributions useful for modelling diet estimates obtained from QFASA (described in Section 1.1) due to the amalgamation property described in Property 2.4. This property allows marginal distributions to be obtained. This is useful for FA analysis since multivariate inference such as confidence regions is often impossible due to the high dimensionality of the data, and small sample sizes. Therefore, univariate inference based on these marginal distributions is much more practical.

A new class of distributions was extended onto the simplex by Mateu-Figueras et al. (2005) which was based on the skew-normal distributions introduced by Azzalini

& Valle (1996). This class of distributions is an extension of the additive logistic distribution that allows for low to moderate levels of skewness in the transformed data. The original skew-normal distribution was defined as follows:

Definition 2.23. *A D-dimensional random vector* $\boldsymbol{Y}$ *follows a multivariate skew-normal distribution* $\mathcal{SN}^D(\boldsymbol{\mu}, \boldsymbol{\Sigma}, \boldsymbol{\alpha})$*, if it is continuous with density:*

$$\begin{aligned} f(\boldsymbol{y}; \boldsymbol{\mu}, \boldsymbol{\Sigma}, \boldsymbol{\alpha}) &= 2MVN^D(\boldsymbol{y}; \boldsymbol{\mu}, \boldsymbol{\Sigma})\phi(\boldsymbol{\alpha}^T\boldsymbol{\Omega}^{-1}(\boldsymbol{y}-\boldsymbol{\mu})) \\ &= \frac{2}{(2\pi)^{D/2}|\boldsymbol{\Sigma}|^{1/2}} exp\left[-\frac{1}{2}(\boldsymbol{y}-\boldsymbol{\mu})^T\boldsymbol{\Sigma}^{-1}(\boldsymbol{y}-\boldsymbol{\mu})\right]\phi(\boldsymbol{\alpha}^T\boldsymbol{\Omega}^{-1}(\boldsymbol{y}-\boldsymbol{\mu})) \end{aligned}$$

where $MVN^D(\boldsymbol{y}; \boldsymbol{\mu}, \boldsymbol{\Sigma})$ *is the density function of the D-dimensional normal distribution evaluated at* $\boldsymbol{y}$*, with mean* $\boldsymbol{\mu}$ *and variance-covariance matrix* $\boldsymbol{\Sigma}$*,* $\phi(\cdot)$ *is the* $N(0,1)$ *distribution function,* $\boldsymbol{\Omega}$ *is the square root of the diagonal matrix with standard deviations of the* $\boldsymbol{\Sigma}$ *diagonal, and* $\boldsymbol{\alpha}$ *is a D-dimensional shape parameter.*

Note: When $\boldsymbol{\alpha} = \boldsymbol{0}$*,* $\boldsymbol{Y}$ *is distributed as* $MVN^D(\boldsymbol{\mu}, \boldsymbol{\Sigma})$*.* $\boldsymbol{\alpha}$ *determines the shape of the distribution and the direction of maximum skewness.*

Mateu-Figueras et al. (2005) extended this distribution onto the simplex with the additive logistic transformation, and called it the additive logistic skew-normal distribution defined below.

Definition 2.24. *A D-dimensional composition* $\boldsymbol{U}_o$ *is said to have an additive logistic skew-normal distribution,* $\mathcal{LS}^d(\boldsymbol{\mu}_a, \boldsymbol{\Sigma}_a, \boldsymbol{\alpha})$*, when* $\boldsymbol{U}_a = alr(\boldsymbol{U}_o)$ *has a skew-normal distribution,* $\mathcal{SN}^d(\boldsymbol{\mu}_a, \boldsymbol{\Sigma}_a)$*. That is, if* $\boldsymbol{U}_o$ *has density:*

$$\begin{aligned} f(\boldsymbol{u}_o; \boldsymbol{\mu}_a, \boldsymbol{\Sigma}_a, \boldsymbol{\alpha}) = \frac{2}{(2\pi)^{d/2}|\boldsymbol{\Sigma}_a|^{1/2}(\prod_{i=1}^D u_{oi})} exp\left[-\frac{1}{2}(\boldsymbol{u}_a-\boldsymbol{\mu}_a)^T\boldsymbol{\Sigma}_a^{-1}(\boldsymbol{u}_a-\boldsymbol{\mu}_a)\right] \times \\ \phi(\boldsymbol{\alpha}^T\boldsymbol{\Omega}^{-1}(\boldsymbol{y}-\boldsymbol{\mu}_a)) \end{aligned}$$

Note: $\frac{1}{(\prod_{i=1}^D u_{oi})}$ *is the jacobian of the alr transformation and gets included in the density after applying the change of variables method to the density in Definition 2.23.*

Similarly, Stewart (2005) defined the multiplicative logistic skew-normal distribution below.

Definition 2.25. *A D-dimensional composition $\boldsymbol{U}_o$ is said to have a multiplicative logistic skew-normal distribution, $\mathcal{MS}^d(\boldsymbol{\mu}_m, \boldsymbol{\Sigma}_m, \boldsymbol{\alpha})$, when $\boldsymbol{U}_m$, the multiplicative logistic transformation of $\boldsymbol{U}_o$, defined in Definition 2.22, has a skew-normal distribution, $\mathcal{SN}^d(\boldsymbol{\mu}_m, \boldsymbol{\Sigma}_m)$.*

The distributions presented above can be used to model the diet estimates obtained from QFASA. However, as discussed in Section 2.2.2, sometimes essential zeros are present in the diets; that is, there is a true absence of a prey species in the diet of the predator. In this case, we cannot use the imputation techniques proposed earlier to replace rounded zeros, so we must model the zeros as well. Several zero-inflated distributions have been proposed to do just that.
The three zero-inflated distributions described below refer to marginal distributions. The first mixture model relies on the multiplicative logistic normal distribution.

Definition 2.26. *The simplified probability distribution for $SMix\mathcal{M}(\theta_k, \mu_k, \sigma_k^2)$ proposed in Stewart & Field (2011), is given by:*

$$f_k(p_k) = \begin{cases} \theta_k & \text{if } p_k = 0 \\ (1-\theta_k)\mathcal{M}(\mu_k, \sigma_k^2) & \text{if } 0 < p_k < 1 \\ 0 & \text{otherwise} \end{cases}$$

where p_k is the k^{th} component of the composition, $f_k(\cdot)$ refers to the probability distribution of the k^{th} component of the composition, and $\mathcal{M}(\mu_k, \sigma_k^2)$ is the multiplicative logistic distribution with mean μ_k and variance σ_k^2.

A natural extension of this distribution is to include a parameter allowing for skewness.

Definition 2.27. *Stewart & Field (2011) also proposed a mixture model that is based on the multiplicative logistic skew-normal distribution. The probability distribution*

for $SMixS_k\mathcal{M}(\theta_k, \mu_k, \sigma_k^2, \alpha_k)$ *is given by:*

$$f_k(p_k) = \begin{cases} \theta_k & \text{if } p_k = 0 \\ (1-\theta_k)S_k\mathcal{M}(\mu_k, \sigma_k^2) & \text{if } 0 < p_k < 1 \\ 0 & \text{otherwise} \end{cases}$$

where p_k *is the* k^{th} *component of the composition,* $f_k(\cdot)$ *refers to the probability distribution of the* k^{th} *component of the composition, and* $S_k\mathcal{M}(\mu_k, \sigma_k^2, \alpha_k)$ *is the multiplicative logistic distribution with mean* μ_k*, variance* σ_k^2 *and skew parameter* α_k*.*

A similar model is discussed in Stewart (2013) using the beta distribution. The beta distribution is a natural choice when dealing with proportions since a beta-distributed random variable has a range from 0 to 1. In order to include zeros, a zero-inflated model was proposed.

Definition 2.28. *The zero-inflated beta distribution is defined as:*

$$f_k(p_k) = \begin{cases} \theta_k & \text{if } p_k = 0 \\ (1-\theta_k)\beta(\mu_k, \phi_k) & \text{if } 0 < p_k < 1 \\ 0 & \text{otherwise} \end{cases}$$

where p_k *is the* k^{th} *component of the composition,* $f_k(\cdot)$ *refers to the probability distribution of the* k^{th} *component of the composition, and* $\beta(\mu_k, \phi_k)$ *is the beta distribution with rate* μ_k *and scale* ϕ_k*.*

Recently, Tsagris & Stewart (2018b) proposed an α-folded multivariate normal distribution for compositional data which allows a multivariate distribution to be fit on $\mathcal{S}^d$ through the parameter α. It uses the α-folded transformation from $\mathbf{y} \in \mathbb{R}^d$ to $\mathbf{u}_o \in \mathcal{S}^d$ described below:

$$\mathbf{u}_o = \begin{cases} g_0^\alpha(\mathbf{y}) & \text{if } \mathbf{y} \in \mathbb{A}^d \\ g_1^\alpha(\mathbf{y}) & \text{if } \mathbf{y} \in \mathbb{R}^d \setminus \mathbb{A}^d \end{cases}$$

where $g_0^\alpha(\mathbf{y}) = \mathbf{k}_\alpha^{-1}(\mathbf{H}^T\mathbf{y})$, $\mathbf{H}$ is the Helmert Matrix (a square orthogonal matrix; Lancaster (1965)), $g_1^\alpha(\mathbf{y}) = \mathbf{k}_\alpha^{-1}\left(\frac{\mathbf{H}^T\mathbf{y}}{q_\alpha^{*2}(\mathbf{y})}\right)$, $\mathbb{A}_\alpha^d = \left\{\mathbf{H}\mathbf{k}_\alpha | -\frac{1}{\alpha} \leq k_{i,\alpha} \leq \frac{d}{\alpha}, \sum_{i=1}^{d+1} k_{i,\alpha} = 0\right\}$,

and the functions involved are as follows:

$$\mathbf{k}_\alpha^{-1}(\mathbf{m}) = \frac{(1+\alpha m_i)^{1/\alpha}}{\sum_{j=1}^{D}(1+\alpha m_i)^{1/\alpha}}, \text{for } i = 1, ..., D$$

$$q_\alpha^*(\mathbf{y}) = \alpha \min\{\mathbf{H}^T\mathbf{Y}\}.$$

If $\mathbf{Y}$ shown above follows a multivariate normal distribution, $\mathbf{U}_o$ is said to have an α-folded multivariate normal distribution.

2.5 Measures of Location and Spread

Aitchison & Bacon-Shone (1999) constructed several approximations that are useful for a maximum likelihood approach which are described in Section 3.2. These were constructed to resolve the issue that the distribution of a convex linear combination of d-dimensional additive logistic normal random compositions is not easily determined. In these approximations, Aitchison uses the reparameterization where the centre (the theoretical closed geometric mean) is defined on the untransformed scale and the variation matrix defined in Aitchison (1986) is used in place of a covariance matrix. These are the arguments used to describe the additive logistic normal distribution $\mathcal{L}(\boldsymbol{\xi}, \mathbf{T})$, which says that the alr transformation follows a normal distribution with the untransformed composition having theoretical closed geometric mean $\boldsymbol{\xi}$ and variation matrix $\mathbf{T}$. Note, this is different notation than used in this thesis (see Definition 2.21). Therefore, let's first define these measures of centre and variability, and then relate them back to our notation for use throughout the rest of the thesis.

Definition 2.29. *The theoretical closed geometric mean of a composition $\boldsymbol{\xi}$, also called Aitchison's mean, is defined in Aitchison (1986) as:*

$$\boldsymbol{\xi} = \mathcal{C}[g_1, ..., g_D]$$

where g_j refers to the theoretical geometric mean of the j^{th} components, $\log g_j = \int \log(u_{oj}) f(u_{oj}) du_{oj}$. Here $f(u_{oj})$ represents the probability density function of u_{oj}.

Definition 2.30. *The empirical form of Definition 2.29 (Aitchison's mean) can be found by using the following formula:*

$$mean_A(\boldsymbol{u}_o) = \mathcal{C}[\hat{g}_1, ..., \hat{g}_D]$$

where $\hat{g}_j$ *refers to the geometric mean of the* j^{th} *component,* $\hat{g}_j = (\prod_{i=1}^{n} u_{oji})^{1/n}$.

Definition 2.31. *The variation matrix,* $\boldsymbol{T} = [\tau_{ij}]$*, is defined in Aitchison (1986) as:*

$$\tau_{ij} = var\left(\log \frac{u_{oi}}{u_{oj}}\right), i, j = 1, ..., D.$$

With these parameterizations, Aitchison & Bacon-Shone (1999) then say that $\mathcal{L}^d(\boldsymbol{\xi}, \mathbf{T})$ represents a logistic normal distribution with centre $\boldsymbol{\xi}$ and variation matrix $\mathbf{T}$.

We can relate these new parameters to the mean and variance-covariance matrix of the alr transformation, $\boldsymbol{\mu}_{\mathrm{a}}$ and $\boldsymbol{\Sigma}_{\mathrm{a}}$, by:

$$\mu_{\mathrm{a}j} = \log \frac{\xi_j}{\xi_D} = \mathrm{alr}(\boldsymbol{\xi})_j \qquad j = 1, ..., d \tag{2.3}$$

$$\sigma_{\mathrm{a}ij} = \frac{1}{2}(\tau_{iD} + \tau_{jD} - \tau_{ij}) \qquad i, j = 1, ..., D \tag{2.4}$$

In order to keep notation clear, the transformations, their means, covariance matrices, and space are organized in Table 2.1.

Transformation	Notation	Space	Mean	Covariance Matrix
No transformation	$\mathbf{u}_{\mathrm{o}}$	$\mathcal{S}^D$	$\boldsymbol{\xi}$	$\mathbf{T} = [\tau_{ij}]$
Alr transformation	$\mathbf{u}_{\mathrm{a}} = \mathrm{alr}(\mathbf{u}_{\mathrm{o}})$	$\mathbb{R}^d$	$\boldsymbol{\mu}_{\mathrm{a}}$	$\boldsymbol{\Sigma}_{\mathrm{a}} = [\sigma_{\mathrm{a}ij}]$
Clr transformation	$\mathbf{u}_{\mathrm{c}} = \mathrm{clr}(\mathbf{u}_{\mathrm{o}})$	$\mathbb{R}^D$	$\boldsymbol{\mu}_{\mathrm{c}}$	$\boldsymbol{\Sigma}_{\mathrm{c}} = [\sigma_{\mathrm{c}ij}]$
Ilr transformation	$\mathbf{u} = \mathrm{ilr}(\mathbf{u}_{\mathrm{o}})$	$\mathbb{R}^d$	$\boldsymbol{\mu}$	$\boldsymbol{\Sigma} = [\sigma_{ij}]$

Table 2.1: Notation for transformations between the simplex and real space.

If the mean is known for a certain transformation of a composition, we can relate that back to Aitchison's mean of the original composition $\mathbf{u}_{\mathrm{o}}$ using the following formula from Aitchison (2005):

$$\boldsymbol{\xi} = \text{mean}_A(\mathbf{u}_\text{o}) = \text{alr}^{-1}(\boldsymbol{\mu}_\text{a}) = \text{alr}^{-1}(\text{mean}[\text{alr}(\mathbf{u}_\text{o})]) \tag{2.5}$$

where mean here represents the arithmetic mean. The relationships between transformations are given in Egozcue et al. (2003) and are shown to simply be matrix multiplications between the transformations. This, combined with properties of expected values, leads us to extend Equation 2.5 to obtain:

$$\text{mean}_A(\mathbf{u}_\text{o}) = *\text{lr}^{-1}(\text{mean}[*\text{lr}(\mathbf{u}_\text{o})]) \tag{2.6}$$

where $*$lr represents any of the transformations alr, clr or ilr.

Similarly, we can convert the covariance matrix of one transformed composition to another using the set of equations shown below from Aitchison (1986). In the equations, the following shorthands are used:

$$\tau_{i\cdot} = \frac{1}{D}\sum_{j=1}^{D}\tau_{ij}, \qquad \tau_{\cdot\cdot} = \frac{1}{D^2}\sum_{i=1}^{D}\sum_{j=1}^{D}\tau_{ij}$$

$$\sigma_{\text{a}i\cdot} = \frac{1}{D}\sum_{j=1}^{D}\sigma_{\text{a}ij}, \qquad \sigma_{\text{a}\cdot\cdot} = \frac{1}{D^2}\sum_{i=1}^{D}\sum_{j=1}^{D}\sigma_{\text{a}ij}$$

The equations below show the conversions of the elements of the covariance matrices from one transformation to another.

$$\begin{aligned}
\mathbf{T} \to \boldsymbol{\Sigma}_\text{a}: &\qquad \sigma_{ij} = \frac{1}{2}(\tau_{iD} + \tau_{jD} - \tau_{ij}) \\
\mathbf{T} \to \boldsymbol{\Sigma}_\text{c}: &\qquad \sigma_{\text{c}ij} = \frac{1}{2}(\tau_{i\cdot} + \tau_{j\cdot} - \tau_{ij} - \tau_{\cdot\cdot}) \\
\boldsymbol{\Sigma}_\text{a} \to \mathbf{T}: &\qquad \tau_{ij} = \sigma_{\text{a}ii} + \sigma_{\text{a}jj} - 2\sigma_{\text{a}ij} \\
\boldsymbol{\Sigma}_\text{a} \to \boldsymbol{\Sigma}_\text{c}: &\qquad \sigma_{\text{c}ij} = \sigma_{\text{a}ij} - \sigma_{\text{a}i\cdot} - \sigma_{\text{a}j\cdot} + \sigma_{\text{a}\cdot\cdot} \\
\boldsymbol{\Sigma}_\text{c} \to \mathbf{T}: &\qquad \tau_{ij} = \sigma_{\text{c}ii} + \sigma_{\text{c}jj} - 2\sigma_{\text{c}ij} \\
\boldsymbol{\Sigma}_\text{c} \to \boldsymbol{\Sigma}_\text{a}: &\qquad \sigma_{\text{a}ij} = \sigma_{\text{c}ij} - \sigma_{\text{c}iD} - \sigma_{\text{c}jD} + \sigma_{\text{c}DD}
\end{aligned} \tag{2.7}$$

The relationships above can also be depicted using matrix algebra, in lieu of element by element calculations. These are shown below.

$$
\begin{aligned}
\mathbf{T} \to \mathbf{\Sigma}_{\mathrm{a}}: &\qquad \mathbf{\Sigma}_{\mathrm{a}} = -\frac{1}{2}\mathbf{F}\mathbf{T}\mathbf{F}^T \\
\mathbf{T} \to \mathbf{\Sigma}_{\mathrm{c}}: &\qquad \mathbf{\Sigma}_{\mathrm{c}} = -\frac{1}{2}\mathbf{G}\mathbf{T}\mathbf{G} \\
\mathbf{\Sigma}_{\mathrm{a}} \to \mathbf{T}: &\qquad \mathbf{T} = \text{First convert to } \mathbf{\Sigma}_{\mathrm{c}} \text{ then convert to } \mathbf{T} \\
\mathbf{\Sigma}_{\mathrm{a}} \to \mathbf{\Sigma}_{\mathrm{c}}: &\qquad \mathbf{\Sigma}_{\mathrm{c}} = \mathbf{F}^T\mathbf{H}^{-1}\mathbf{\Sigma}_{\mathrm{a}}\mathbf{H}^{-1}\mathbf{F} \\
\mathbf{\Sigma}_{\mathrm{c}} \to \mathbf{T}: &\qquad \mathbf{T} = \mathbf{J}\mathrm{diag}(\mathbf{\Sigma}_{\mathrm{c}}) + \mathrm{diag}(\mathbf{\Sigma}_{\mathrm{c}})\mathbf{J} - 2\mathbf{\Sigma}_{\mathrm{c}} \\
\mathbf{\Sigma}_{\mathrm{c}} \to \mathbf{\Sigma}_{\mathrm{a}}: &\qquad \mathbf{\Sigma}_{\mathrm{a}} = \mathbf{F}\mathbf{\Sigma}_{\mathrm{c}}\mathbf{F}^T
\end{aligned}
\tag{2.8}
$$

where $\mathbf{J}_d$ is a matrix of units, $\mathbf{F}_{d,D} = [\mathbf{I}_d, \mathbf{j}_d]$, $\mathbf{I}_d$ is the d-dimensional identity matrix, $\mathbf{j}_d$ is a column vector of units, $\mathbf{G}_D = \mathbf{I}_D - D^{-1}\mathbf{J}_D$, $\mathbf{H}_d = \mathbf{I}_d + \mathbf{J}_d$, and $\mathrm{diag}(\mathbf{\Sigma}_c)$ is a diagonal matrix with the diagonal entries equal to the diagonal entries of $\mathbf{\Sigma}_c$.

To include the covariance matrix after an ilr transformation into these equations, which we have denoted $\mathbf{\Sigma}$, we first need to establish several relationships. Let $\mathbf{U}_o$ be a random composition of length D, then from Egozcue et al. (2003), we know that:

$$
\begin{aligned}
\mathrm{ilr}(\mathbf{U}_o) = \mathbf{U} &= \mathrm{clr}(\mathbf{U}_o)\mathbf{V} = \mathbf{U}_c\mathbf{V} \\
\mathrm{clr}(\mathbf{U}_o) = \mathbf{U}_c &= \mathrm{ilr}(\mathbf{U}_o)\mathbf{V}^T = \mathbf{U}\mathbf{V}^T \\
\mathbf{V}^T\mathbf{V} &= \mathbf{I}_{D-1}
\end{aligned}
\tag{2.9}
$$

From above, we can find the relationships between the variance-covariance matrices of the ilr transformation and clr transformation as follows:

$$
\begin{aligned}
\mathbf{U}_c &= \mathbf{U}\mathbf{V}^T \\
\mathbf{U}_c^T &= [\mathbf{U}\mathbf{V}^T]^T \\
\mathbf{U}_c^T &= \mathbf{V}\mathbf{U}^T \\
\mathrm{Var}(\mathbf{U}_c^T) &= \mathrm{Var}(\mathbf{V}\mathbf{U}^T) \\
\mathbf{\Sigma}_c &= \mathbf{V}\mathbf{\Sigma}\mathbf{V}^T
\end{aligned}
\tag{2.10}
$$

Using this last equality, and using the relationship between $\mathbf{V}$ and the identity matrix shown above, we get:

$$\begin{aligned}
\boldsymbol{\Sigma}_c &= \mathbf{V}\boldsymbol{\Sigma}\mathbf{V}^T \\
\mathbf{V}^T\boldsymbol{\Sigma}_c\mathbf{V} &= \mathbf{V}^T\mathbf{V}\boldsymbol{\Sigma}\mathbf{V}^T\mathbf{V} \\
\mathbf{V}^T\boldsymbol{\Sigma}_c\mathbf{V} &= \mathbf{I}_{D-1}\boldsymbol{\Sigma}\mathbf{I}_{D-1} \\
\mathbf{V}^T\boldsymbol{\Sigma}_c\mathbf{V} &= \boldsymbol{\Sigma}
\end{aligned} \tag{2.11}$$

Thus, the two relationships between $\boldsymbol{\Sigma}$ and $\boldsymbol{\Sigma}_c$ can be added to the list above:

$$\begin{aligned}
\boldsymbol{\Sigma} \to \boldsymbol{\Sigma}_c : &\qquad \mathbf{V}^T\boldsymbol{\Sigma}_c\mathbf{V} \\
\boldsymbol{\Sigma}_c \to \boldsymbol{\Sigma} : &\qquad \mathbf{V}\boldsymbol{\Sigma}\mathbf{V}^T
\end{aligned} \tag{2.12}$$

2.6 Convex Linear Combinations of Compositions

In Aitchison & Bacon-Shone (1999), three approximations are used to obtain the mean and variation matrix of a convex linear combination of compositions on the untransformed scale. In the approximations below, if u_o is said to follow $\mathcal{L}^d(\boldsymbol{\xi}, \mathbf{T})$, it means that $\mathbf{U}_a = alr(\mathbf{U}_o)$ follows $\text{MVN}^{D-1}(\boldsymbol{\mu}_a, \boldsymbol{\Sigma}_a)$, where $\boldsymbol{\mu}_a = \text{alr}(\boldsymbol{\xi})$ as seen in Equation 2.6 and $\boldsymbol{\Sigma}_a = -\frac{1}{2}\mathbf{F}\mathbf{T}\mathbf{F}^T$ as seen in Equation 2.7. These approximations are described below.

Approximation 2.1. *Let $\boldsymbol{Y}$ be a convex linear combination of compositions, $\boldsymbol{Y} = \pi_1\boldsymbol{U}_{o1} + \pi_2\boldsymbol{U}_{o2} + ... + \pi_C\boldsymbol{U}_{oC}$, where $\boldsymbol{U}_{o1}$, ..., $\boldsymbol{U}_{oc}$ are independently distributed as $\mathcal{L}^d(\boldsymbol{\xi}_1, \boldsymbol{T}_1)$, $\mathcal{L}^d(\boldsymbol{\xi}_2, \boldsymbol{T}_2)$, ..., $\mathcal{L}^d(\boldsymbol{\xi}_C, \boldsymbol{T}_C)$. Then $\boldsymbol{Y}$ is approximately distributed as $\mathcal{L}^d(\boldsymbol{\eta}_o, \boldsymbol{\Theta})$, $\boldsymbol{\Theta} = [\theta_{ij}]$, and*

$$\boldsymbol{\eta}_o = \sum_{b=1}^{C} \pi_b \boldsymbol{\xi}_b$$

$$\theta_{ij} = -\frac{1}{2}\sum_{b=1}^{C}\sum_{k=1}^{D}\sum_{l=1}^{D} G_{bijk}G_{bijl}\tau_{bkl}$$

where,

$$G_{bijk} = \rho_{bi}(\delta_{ik} - \xi_{bk}) - \rho_{bj}(\delta_{jk} - \xi_{bk}), \qquad \rho_{bi} = \pi_b\xi_{bi}/\eta_{oi},$$

and δ_{ik} is the Kronecker delta, equal to 1 when $k = i$, and equal to 0 when $k \neq i$.

This approximation is derived from the first-order Taylor expansion of the log ratio of Y_i and Y_j. Also note that if the mixture places full weight onto one composition, that composition's original mean and variation matrix are obtained for $\boldsymbol{\eta}_o$ and $\boldsymbol{\Theta}$, respectively.

The second approximation considers the case where the mixing vector $\boldsymbol{\pi}$ is not fixed, but varies compositionally according to a $\mathcal{L}^{C-1}(\boldsymbol{\alpha}, \boldsymbol{\Omega})$ distribution.

Approximation 2.2. *Let $\boldsymbol{Y}$ be a convex linear combination of compositions, $\boldsymbol{Y} = \pi_1 \boldsymbol{U}_{o1} + \pi_2 \boldsymbol{U}_{o2} + ... + \pi_C \boldsymbol{U}_{oC}$, where $\boldsymbol{U}_{o1}$, ..., $\boldsymbol{U}_{oC}$ are independently distributed as $\mathcal{L}^d(\boldsymbol{\xi}_1, \boldsymbol{T}_1)$, $\mathcal{L}^d(\boldsymbol{\xi}_2, \boldsymbol{T}_2)$, ..., $\mathcal{L}^d(\boldsymbol{\xi}_C, \boldsymbol{T}_C)$, and $\boldsymbol{\pi}$ is distributed as $\mathcal{L}^{C-1}(\boldsymbol{\alpha}, \boldsymbol{\Omega})$. Then $\boldsymbol{Y}$ is approximately distributed as $\mathcal{L}^d(\boldsymbol{\kappa}, \boldsymbol{\Lambda})$, $\boldsymbol{\Lambda} = [\lambda_{ij}]$, where*

$$\boldsymbol{\kappa} = \sum_{b=1}^{C} \alpha_b \boldsymbol{\xi}_b$$

$$\lambda_{ij} = -\frac{1}{2} \sum_{b=1}^{C} \sum_{k=1}^{D} \sum_{l=1}^{D} H_{bijk} H_{bijl} \tau_{bkl} - \frac{1}{2} \sum_{a=1}^{C} \sum_{b=1}^{C} B_{aij} B_{bij} \omega_{ab}$$

where

$$H_{bijk} = \chi_{bi}(\delta_{ik} - \xi_{bk}) - \chi_{bj}(\delta_{jk} - \xi_{bk}), \qquad \chi_{bi} = \alpha_b \xi_{bi} / \kappa_i, \qquad B_{bij} = \chi_{bi} - \chi_{bj}$$

Note that Approximation 1 is the special case of $\boldsymbol{\Omega} = 0$ and $\boldsymbol{\alpha} = \boldsymbol{\pi}$. A third approximation is introduced where, instead of including a varying mixture, the mixture is considered fixed, and a perturbation effect is included.

Approximation 2.3. *Let's refer to the convex linear combination $\pi_1 \boldsymbol{U}_{o1} + \pi_2 \boldsymbol{U}_{o2} + ... + \pi_c \boldsymbol{U}_{oc}$ as $cvx(\boldsymbol{U}_o, \boldsymbol{\pi})$. Then let $\boldsymbol{Y} = cvx(\boldsymbol{U}_o, \boldsymbol{\pi}) \circ \boldsymbol{X}$, where $\boldsymbol{X}$ is a compositional perturbation described in Aitchison (1986), distributed as $\mathcal{L}^d(\boldsymbol{e}, \boldsymbol{\Psi})$, where $\boldsymbol{e} = \frac{1}{D}(1, 1, ..., 1)$ is the identity of the perturbation group. Then $\boldsymbol{Y}$ is approximately distributed as $\mathcal{L}^d(\boldsymbol{\eta}_o, \boldsymbol{\Theta} + \boldsymbol{\Psi})$, where $\boldsymbol{\eta}_o$ and $\boldsymbol{\Theta}$ are given in Approximation 1.*

These three models are referred to as the fixed-mixture model, the convolution model, and the perturbation model, respectively. For our methods, we will work with the perturbation model. Aitchison & Bacon-Shone (1999) performed simulations to ensure that the approximations yield accurate results, and they found that the parameter estimates are within appropriate practical values and the normality appears satisfied for most practical situations.

2.7 Inference for Compositions

2.7.1 Testing for Difference in Diet

Stewart et al. (2014) proposed that testing for a difference in location of FA signatures is a good indicator for a difference in diet, as changes in diet are reflected in the FA signatures. While it could be argued that a difference in FAs could be due to differences in metabolism between individuals, all existing methods for quantitative diet estimation treat metabolic rates for one species as a constant. Therefore, the premise that differences in FA signatures is a good indicator for differences in diet is used in the next three tests for comparing diet.

Two Independent Samples

Suppose two independent samples of predators are collected of size n_1 and n_2 respectively, each of dimension m. Stewart et al. (2014) introduced a test based on the nonparametric permutation test for comparing two means, in order to compare the mean FA signatures of the two groups. Let $\mathbf{y}_{1i}$ be the i^{th} predator's FA signature in the first group, and $\mathbf{y}_{2j}$ be the j^{th} predator's FA signature in the second group, then the test statistic is as follows:

$$T = \sum_{i_1}^{n_1} \sum_{i_2}^{n_2} \text{dist}(\mathbf{y}_{1i_1}, \mathbf{y}_{2i_2}) \tag{2.13}$$

where "dist" denotes either Aitchison's distance defined in Definition 2.15, or chi-square distance defined in Definition 2.19.

This test statistic is the used to perform the multivariate permutation test described

below. The steps to this test are as follows:

1. Compute the test statistic T, defined in Equation 2.13, between $\mathbf{y}_1$ and $\mathbf{y}_2$.
2. for r = 1,...,R
   (a) Pool the two samples.
   (b) Permute the $n_1 + n_2$ observations to obtain $\mathbf{y}_i^{*r}$, $i = 1, ..., n_1 + n_2$.
   (c) Let $\mathbf{y}_{1i}^{*r} = \mathbf{y}_i^{*r}$, $i = 1, ..., n_1$, and $\mathbf{y}_{2i}^{*r} = \mathbf{y}_i^{*r}$, $i = n_1 + 1, ..., n_1 + n_2$.
   (d) Compute the test statistic, T^{*r}, defined in Equation 2.13, between $\mathbf{y}_1^{*r}$ and $\mathbf{y}_2^{*r}$.
3. Compute the p-value.

$$p^{MPT} = \frac{\#\{T^{*r} \geq T\}}{R}$$

2.7.2 Paired Samples

Stewart et al. (2014) also introduced a test for comparing FA signatures in paired samples, based on the univariate matched pair randomization p-value in Davison & Hinkley (1997). Similar to the independent case, the test statistic is based on a distance measure, which could be one of the two measures described below.

The first distance measure treats the zeros as rounded zeros. The zeros are modified, the data is log-ratio transformed, and then the distance is calculated between the before and after signatures for each individual. Specifically, it is calculated as:

$$\mathbf{d}_i^{log} = \mathbf{u}_{Aai} - \mathbf{u}_{Bai}$$

where $\mathbf{u}_{Aa} = alr(\mathbf{y}_B)$ and $\mathbf{u}_{Ba} = alr(\mathbf{y}_A)$, and $\mathbf{y}_B$ and $\mathbf{y}_A$ are the "Before" and "After" FA signatures, respectively.

Alternatively, a measure based on the chi-square distance described in definition 2.19 can be computed as:

$$\mathbf{d}_i^{CS} = \frac{1}{\gamma}\sqrt{2m}\mathbf{C}^{-1/2}\left(\mathbf{z}_{Ai}^{CS} - \mathbf{z}_{Bi}^{CS}\right)$$

where $\mathbf{z}_A^{CS}$ and $\mathbf{z}_{Bi}^{CS}$ are the re-closed power transformed data and $\mathbf{C}$ is a m x m diagonal matrix with diagonal elements $c_{ij} = \mathbf{u}_{Aai} - \mathbf{u}_{Bai}$ when one of $\mathbf{u}_{Aai}$ and $\mathbf{u}_{Bai}$ are non zero, and $c_{ij} = 1$ when both are zero. The choice of γ is described in detail in Stewart et al. (2014). The distances of the before and after FA signatures are compared to the distances between the two-part corresponding subcompositions when the data are power-transformed.

One of the two distance measures described above is selected, and used for the computation of the test statistic, which is calculated as follows:

$$T = \left(\sum_{j=1}^{m} \bar{d}_j^2\right)^{1/2} \tag{2.14}$$

where $\bar{d}_j = \frac{1}{n}\sum_{i=1}^{n} d_{ij}$, $j = 1, ..., m$, and d_{ij} is the difference, computed using one of the two methods described above, between the j^{th} FA for the i^{th} individual.

The test is referred to as the multivariate randomization test and can be performed using the following steps:

1. Compute the differences $\mathbf{d}i$, $i = 1, ..., n$.
2. Compute the test statistic shown in Equation 2.14 using the differences in 1.
3. for r = 1,...,R
   (a) For the i^{th} observation, randomly select either +1 or -1. and call this s_i^{r*}, $i = 1, ..., n$.
   (b) Compute $\mathbf{d}_i^{*r} = s_i^{r*}\mathbf{d}_i$, $i = 1, ..., n$.

(c) Compute T^{*r} using Equation 2.14 and $\mathbf{d}_i^{*r}$.

4. Compute the p-value.

$$p^{MRT} = \frac{\#\{T^{*r} \geq T\}}{R}$$

2.7.3 More than 2 Independent Groups

Suppose we have more than 2 independent groups, and we want to compare diet proportions using FA signatures. A multivariate ANOVA technique would be required. Although Aitchison (1986) proposes the possibility of completing a MANOVA on compositions by transforming the compositions using log-ratios, this method is only applicable if all of the assumptions of a MANOVA are valid. However the traditional MANOVA has strong assumptions that frequently do not agree with compositional data. These restrictions include the assumption that the transformed data is approximately multivariate normal, the appropriate use of Euclidean distances, and that the sample size be larger than the number of variables. With compositional data however, sometimes due to zeros, a transformation is not possible or the transformed data may not follow a multivariate normal distribution. More importantly, often the number of variables is larger than the sample size, as is the case with most FA signatures. Also, as discussed, Euclidean distances are not subcompositionally coherent, making them inappropriate measures for compositions.

In Steeves et al. (2016), a non-parametric MANOVA based on McArdle & Anderson (2001) was used to detect differences in diet proportions based on testing FA signatures. This test is appropriate for compositional data since not only are the assumptions of log-normality, Euclidean distances, and sample size unnecessary, but it also allows use of the distance between compositions to carry out a MANOVA-type test. Similar to a parametric MANOVA, it requires $\mathbf{Y}$, a sample of n units in p variables, as well as a $n \times k$ model matrix or design matrix $\mathbf{X}$. The goal is to test the hypothesis that the model parameters have no effect. That is, $H_o : \boldsymbol{\beta} = \mathbf{0}$ in the equation $\mathbf{Y} = \mathbf{X}\boldsymbol{\beta} + \boldsymbol{\epsilon}$. Instead of the traditional parametric approach, we base our test statistic on a distance matrix, or semi-metric distance matrix, $\boldsymbol{\Delta} = [d_{ij}]$ of the sample of FA signatures, $\mathbf{Y}$. Using this, we can obtain the test statistic given by:

$$F = \frac{tr(\mathbf{HGH})/(p-1)}{tr[(\mathbf{I}-\mathbf{H})\mathbf{G}(\mathbf{I}-\mathbf{H})]/(n-p)} \tag{2.15}$$

where $\mathbf{H}$ is the hat matrix defined by $\mathbf{H} = \mathbf{X}(\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T$ and $\mathbf{G}$ is Gower's centred matrix defined by $\mathbf{G} = (\mathbf{I} - \frac{1}{n}\mathbf{1}\mathbf{1}^T)\mathbf{A}(\mathbf{I} - \frac{1}{n}\mathbf{1}\mathbf{1}^T)$. Here, $\mathbf{\Delta} = [d_{ij}]$ is described above, and $\mathbf{A} = [a_{ij}] = [-\frac{1}{2}d_{ij}^2]$.

Because this test statistic does not follow the F distribution like in a parametric MANOVA, a permutation method is required to calculate the p-value. If we randomly permute all the observations and calculate the F statistic for all possible permutations, then the p-value is the proportion of these F-statistics that are equal to or larger than the original F from the data. Often it is impossible or inefficient to calculate all possible permutations, so a sufficient sample of permutations can be used. If the desired level of significance is 0.05, then at least 1000 permutations must be computed, and if the desired level of significance is 0.01, then at least 5000 permutations must be computed (McArdle & Anderson (2001)).

The two way design is a natural extension of this, where a partial F test is used. That is, the F statistic is computed as:

$$F = \frac{(SSR_r - SSR_f)/(df_r - df_f)}{SSR_f/df_f} \tag{2.16}$$

where SSR_f and SSR_r are the residual sum of squares for the full and reduced models respectively, and df_f and df_r are the residual sum of squares degrees of freedom for the full and reduced models. This works out to be $N - p + 1$, where p is the number of parameters in the model. The same method for computing the p-value described earlier can be used here.

2.7.4 Regression for Compositions

Dirichlet Component Regression

In Gueorguieva et al. (2008), a regression method was proposed for compositional data that relies on the Dirichlet distribution, the density of which is described in Definition 2.20. The model for each component of the composition $\mathbf{Y}_i$, describes each $\log(\alpha_j)$, ($\boldsymbol{\alpha}$ is the vector of parameters described in Definition 2.20) as separate linear functions of covariates. In other words, each component, $i = 1, ..., k$, uses a log-link function as follows:

$$\log(\alpha_{ij}) = \boldsymbol{\beta}_i^T \mathbf{z}_j$$

where $\mathbf{z}_j$ are the covariates recorded on the jth individual (j = 1,...,n) and the coefficients $\boldsymbol{\beta}_i$ are estimated using maximum likelihood.

Alternatively, as proposed in Tsagris & Stewart (2018a), the following link function could be used:

$$u_{o1} = \frac{1}{1 + \sum_{k=2}^{D} e^{\mathbf{X}^T \boldsymbol{\beta}_k}}, \qquad u_{oi} = \frac{e^{\mathbf{X}^T \boldsymbol{\beta}_i}}{1 + \sum_{k=2}^{D} e^{\mathbf{X}^T \boldsymbol{\beta}_k}}, \text{for } i = 2, ..., D \tag{2.17}$$

Tsagris & Stewart (2018a) also proposed a Dirichlet regression model capable of handling zeros, based on methods discussed in Stewart & Field (2011). Their method assumes there are B populations corresponding to all subsets of non-zero components in composition $\mathbf{Y}$. Let $\theta_b = P(\mathbf{G} = \mathbf{g}_b)$ be the marginal probability that an observation comes from population b, where $\mathbf{g}_b$ is a vector of 1s and 0s corresponding to population b, and $\sum_{b=1}^{B} \theta_b = 1$. Then if$\mathbf{G}$ denotes the vector indexing the non-zero components of $\mathbf{Y}$, the density of $\mathbf{Y}$ with non-zero components corresponding to population b* is given by:

$$f_{\mathbf{Y}}(\mathbf{y}) = \sum_{b=1}^{B} f_{\mathbf{Y},\mathbf{G}}(\mathbf{y}, \mathbf{g}_b) = f_{\mathbf{Y},\mathbf{G}}(\mathbf{y}, \mathbf{g}_{b^*}) \tag{2.18}$$

where $\mathbf{g}_{b^*}$ is the vector of indices corresponding to the non-zero components of $\mathbf{y}$. So, if $\mathbf{y}_{b^*}$ is of length D_{b^*} and denotes the vector containing the non-zero components of $\mathbf{y}$, and $f_{b^*}(\mathbf{y}_{b^*})$ is the density of $\mathbf{Y}_{b^*}$, then:

$$f_{\mathbf{Y}}(\mathbf{y}) = f_{\mathbf{Y},\mathbf{G}}(\mathbf{y}, \mathbf{g}_{b^*}) = f_{\mathbf{Y}|\mathbf{G}}(\mathbf{y}|\mathbf{g}_{b^*})\theta_{b^*} = f_{b^*}(\mathbf{y}_{b^*})\theta_{b^*} \tag{2.19}$$

For this model, Tsagris & Stewart (2018a) chose to use the Dirichlet density. That is, $\mathbf{Y}_{b^*} \sim Dir(\phi_{b^*}, \mathbf{x}^*_{b^*})$, where the x_is are defined in Equation 2.17.

Chapter 3

Implementation of Compositional Models for FA Data

QFASA was developed to estimate diet proportions of marine predators using FA signatures of predators and their assumed prey. Since the initial creation, there have been many improvements including the addition of calibration coefficients into the model to correct for metabolization that occurs within the predator for certain FAs. Fat content has also been included to account for fattier prey species having larger affects on the diet proportions of the predator. However, several drawbacks exist for QFASA that we hope to improve upon. Firstly, we wanted a method that allows for inference so that we could build confidence bounds or hypothesis tests from the diet estimates of the predators. By assuming parametric distributions for the FA signatures (which QFASA does not do), we are able to use "Template Model Builder" or "TMB" to develop methods that will eventually automatically calculate the standard errors. These standard errors are for the asymptotic case, and allow inference to be performed without any extra work, but currently there is some debate as to the accuracy of these standard errors. For these reasons, we will perform inference based on parametric bootstraps in this thesis. Secondly, the individual prey that are consumed by the predator are not themselves included in the sampled preybase. In order to correct for this, we consider the specific prey that the predator consumed as a random effect, and base the distribution of these random effects on that obtained by the sample prey database.

Similar to QFASA, our model takes a linear transformation of the prey FA signatures (or in our case, random effects), and the diet proportions. Since the data is compositional in nature, we had the choice between taking the linear transformation on the untransformed scale (like in QFASA), or on the ilr transformed scale. Although statistically, the model becomes simpler when the linear transformation is performed on the ilr transformed scale, biologically it does not make as much sense. The FA

signatures would be contributing to the diet of the predators as they are consumed, that is, on the untransformed scale, not in a different scale based on logarithms and ratios. Therefore, biologically, it would make more sense to create our model this way.

In order to explain the enhancements to the original QFASA model, we begin this chapter with a detailed explanation of the QFASA methodology, as well as improvements that have been made since it's original proposal. We will then introduce our new maximum likelihood based model for diet estimation, including all required assumptions. Finally, we will end the chapter with a detailed algorithm to utilise the ML methodology.

3.1 QFASA

Recall from Chapter 1 that QFASA utilises a distance minimization algorithm to estimate diet proportions for marine predators from the FA signatures of prey species. To explain this in detail, suppose we have FA signatures of sampled predators, and a prey database, where samples of prey from species believed to be in the diet of the predators are collected, and their FA signatures are recorded. Let $\boldsymbol{\alpha}$ be a composition where the i^{th} element represents the proportion of prey type i in the predator's diet, $\bar{\mathbf{x}}_{oi}$ be the compositional representative FA signature of prey type i (usually the empirical sample mean) on the original, untransformed scale, and $\mathbf{y}_o$ be the FA signature of the predator on the original, untransformed scale. Then we assume that

$$\mathbf{y}_o \approx \sum_{i=1}^{I} \alpha_i \bar{\mathbf{x}}_{oi}. \tag{3.1}$$

In order to estimate the diet proportions $\boldsymbol{\alpha}$, the distance between the observed FA signature $\mathbf{y}_o$ and the linear combination of $\boldsymbol{\alpha}\bar{\mathbf{x}}_o$ is minimized. This is done using a numerical optimiser in R, currently "solnp", and can be implemented on a variety of distance measures, including those described in Section 2.3. Aitchison's distance, shown in Definition 2.15, is the most popular approach to measure distance between two compositional vectors and has been used in QFASA (Bromaghin et al. (2015)). Aitchison's distance satisfies all the distance criteria presented in Aitchison (1992)

(discussed in Section 2.3), including scale invariance, perturbation invariance and subcompositional coherence, however this measure is not feasible when zeros, essential or rounded, are present. The Kullback-Leibler distance was the original distance measure used in QFASA analysis (Iverson et al. (2004a)) and is defined in Definition 2.18. This measure is neither scale invariant, nor subcompositionally dominant, but Iverson et al. (2004a) found it to be a useful and natural measure of distance between two distributions. However in Bromaghin et al. (2015), Aitchison's and KL distance measures were found to yield similar results, although Aitchison's distance yielded estimates with slightly less bias and similar or better root mean squared error (RMSE).

In order to improve the diet estimates obtained from QFASA, several biological corrections have been introduced. The first corrections are referred to as calibration coefficients. These are used to account for metabolization that occurs within the predator, meaning that certain proportions of FAs in the predator will not exactly match those of the consumed prey. Originally, in order to estimate these coefficients, a feeding experiment would need to be conducted where a predator (a seal in Iverson et al. (2004a)) is fed a specific diet for a set amount of time (Iverson et al. (2004a)). The FA signature of the predator (*pred*) and consumed prey(s) (*diet*) are recorded in this instance and used to estimate the k^{th} calibration coefficient, c_k, by taking the 10% trimmed mean of

$$r_{li}^{k} = pred_{ik}/diet_{lk}$$

where k indexes the FA, i indexes the sampled predator, and l indexes the sampled prey. These calibration coefficients get added into the model in Equation 3.1 by replacing the k^{th} FA in the predator's signature (y_{ok}) with y_{ok}^{*} as:

$$\mathrm{y}_{ok}^{*} = \frac{\mathrm{y}_{ok}/c_k}{\sum_s \mathrm{y}_{os}/c_s} \tag{3.2}$$

Recently, Bromaghin et al. (2017b) proposed a method to simultaneously estimate the calibration coefficients along with the diet compositions. They considered the

calibration coefficients as a transformation onto the predator space, and applied this transformation to the mean prey FA signatures as:

$$\bar{\mathrm{x}}_{ik}^{t} = \frac{c_k \bar{\mathrm{x}}_{ik}}{\sum_m c_m \bar{\mathrm{x}}_{im}} \tag{3.3}$$

where c_k is the calibration factor of the k^{th} FA, and $\bar{\mathrm{x}}_{ik}$ is the mean of the k^{th} FA for prey species i. Now, $\bar{\mathrm{x}}_{ik}^{t}$ is in predator space, so we can model the j^{th} predator signature with the diet proportions α_{ji} as:

$$\mathbf{y}_j = \sum_{i=1}^{I} \alpha_{ji} \bar{\mathbf{x}}_i^t \tag{3.4}$$

To optimize the calibration factors and diet estimates simultaneously, Aitchison's distance between the observed and modelled FA signatures are summed over all predators and then minimized.

The second correction takes fat content into account as fattier prey species will contribute more to the FA signature of the predator than less fatty species. If we have a measure of fat content for each prey species, it is easily incorporated into QFASA. We say that $\mathbf{y} = \sum_{i=1}^{I} p_i \bar{\mathbf{x}}_i$ where $\mathbf{p}$ is the composition and where the i^{th} element represents the estimated proportion of species i in the FA signature of the predator, and

$$\alpha_i = \frac{p_i / f_i}{\sum_s p_s / f_s} \tag{3.5}$$

where α_i is the proportion of species i in the diet of the predator and f_i is the fat content of species i. Adding both calibration coefficients and fat content into the QFASA model has been found to significantly improve the diet estimates of real life data as shown in Iverson et al. (2004a) where estimates were much closer to known true diets with these corrections in place than without.

In addition to calibration coefficients and fat content, using specific FA subsets can

help improve diet estimation. Not all FAs are included in QFASA, but typically two different subsets are considered, which vary from one predator species to another. These are referred to as the extended dietary subset and the dietary subset. The extended dietary subset contains FAs that are influenced by both diet, and biosynthesis, and the dietary subset contain FAs that are influenced by diet only. Originally, when using either subset, the FAs to be included would be extracted from the FA signature, and the signature would be rescaled by the partial sum of the FA signature so that it sums to 1 (Iverson et al. (2004a)). However, Bromaghin et al. (2016) believes rescaling the signatures distorts predator-prey relationships and could lead to a bias in diet estimation. They found that when the partial sums (the sum of the proportions after extracting the subset of FAs) differ significantly between prey types, the FA signature structure distorts, causing an increase in bias. He proposes an "augmented" approach for dealing with the subsets where the signatures are not rescaled, but are augmented with an additional proportion equal to 1 minus the partial sum of the FA signature. If the partial sums are similar among prey types, either approach yields similar results.

3.2 Likelihood Model for Diet Estimation

Building on the work of QFASA, described in Section 3.1, our model involves a linear combination of diet proportions and prey FA signatures performed on the untransformed scale, with several modifications described below. The notation for the components of this model are as follows:

- $\mathbf{y}_{oj}$ represents the observed untransformed D-dimensional fatty acid signature of predator j. $\mathbf{y}_j$ is the ilr transformation of $\mathbf{y}_{oj}$.
- $\mathbf{z}_{oji}$ is the unobserved untransformed D-dimensional fatty acid signature of the i^{th} prey species consumed by the j^{th} predator where $i = 1, ..., I$ and $j = 1, ..., n$. This is considered to be an unobserved random effect. $\mathbf{z}_{ji}$ is the ilr transformation of $\mathbf{z}_{oji}$.
- $\mathbf{x}_{oik}$ is the observed D-dimensional fatty acid signature of the k^{th} sampled prey of species i in the prey database. $\mathbf{x}_{ik}$ is the ilr transformation of $\mathbf{x}_{oik}$. Note, each prey species has varying sample sizes.

- $\boldsymbol{\alpha}_j$ represents the unobserved I-dimensional composition of diet proportions for predator j, where α_{ji}, the i^{th} element of $\boldsymbol{\alpha}_j$ is the proportion of prey species i in the predator j's diet.

- $\boldsymbol{\epsilon}_{oj}$ represents the random error associated with predator j. $\boldsymbol{\epsilon}_j$ is the ilr transformation of $\boldsymbol{\epsilon}_{oj}$.

First, for simplicity, we are considering the likelihood of the j^{th} predator with diet $\boldsymbol{\alpha}_j$, and we will ignore calibration and fat content for now. While the fatty acid signature of the predator is directly related to the fatty acid signatures of the specific fish that it ate, these exact fish are unobserved. These prey fish are therefore considered as random effects, which will have to be integrated out to get the appropriate marginal likelihood.

Here, we can think of the predator's FA signature as being a linear combination of the prey random effects, perturbed by some error term $\boldsymbol{\epsilon}_{oj}$. Perturbation was used here since we are dealing with compositional data, and Aitchison & Bacon-Shone (1999) has approximations for the distribution, mean and variance-covariance matrix of this equation. So, if I is the number of species of prey considered, then:

$$\mathbf{y}_{oj} = \left(\sum_{i=1}^{I} \alpha_{ji} \mathbf{z}_{oji} \right) \circ \boldsymbol{\epsilon}_{oj} \tag{3.6}$$

Note, we are only including one FA signature of each prey type in the summation. We could similarly carry out the analysis replacing this one FA signature with a mean (or measure of centre of your choosing) FA signature of n_{prey} prey FA signatures. The distribution of $\mathbf{Y}_{oj}$ cannot be determined explicitly, but we can use Approximation 3 from Aitchison & Bacon-Shone (1999), described in Section 2.6. Note, the notation used here is not the same as the notation in Aitchison & Bacon-Shone (1999) where the distribution of $\mathbf{U}_o$ was said to be $\mathcal{L}^D(\boldsymbol{\xi}, \mathbf{T})$ if $\mathbf{U}_a = \text{alr}(\mathbf{U}_o) \sim \text{MVN}^{D-1}(\text{alr}(\boldsymbol{\xi}), -\frac{1}{2}\mathbf{FTF}^T)$ (see Section 2.5). Here, we will describe the distributions of ilr transformed compositions, and specify the means on the ilr scale, similar to the MVN notation shown above, but using the ilr transformation. Since we are using the ilr transformation in lieu of the alr transformation, we have extended these approximations to the ilr scale using the Equations 2.6 and 2.7. The approximation then tells us that $\mathbf{Y}_j$ will be

approximately multivariate normal, with mean $\text{ilr}(\boldsymbol{\eta}_{oj})$, where $\boldsymbol{\eta}_{oj}$ is given by:

$$\boldsymbol{\eta}_{oj} = \sum_{i=1}^{I} \alpha_{ji} \boldsymbol{\xi}_i \tag{3.7}$$

and with variance-covariance matrix

$$-\frac{1}{2}\mathbf{V}^T\mathbf{G}(\boldsymbol{\Theta} + \mathbf{T}_\epsilon)\mathbf{G}\mathbf{V} \tag{3.8}$$

where $\boldsymbol{\xi}_i$ is Aitchison's mean, described in Section 2.5, of the untransformed FA signatures for prey species i, $\mathbf{T}_\epsilon$ is the variation matrix of $\boldsymbol{\epsilon}_o$ (Definition 2.31) and $\boldsymbol{\Theta}$ is the variation matrix of the $\mathbf{y}_j$s given by

$$\theta_{ab} = -\frac{1}{2}\sum_{l=1}^{I}\sum_{m=1}^{D}\sum_{n=1}^{D} M_{labm}M_{labn}\tau_{mn}.$$

In the above equations, $\mathbf{T} = [\tau_{ab}]$ is the variation matrix of $\mathbf{z}_{oji}$s, $\mathbf{V}$ is a $D \times d$ orthonormal basis of the clr-plane based on the Helmert matrix defined in Equation 2.1, $\mathbf{G} = \mathbf{I}_D - D^{-1}\mathbf{J}_D$, $\mathbf{I}_D$ is the D-dimensional identity matrix, $\mathbf{J}_D$ is a D-dimensional matrix of 1s, and M is defined as

$$M_{labn} = \rho_{la}(\delta_{an} - \xi_{ln}) - \rho_{lb}(\delta_{bn} - \xi_{ln}), \qquad \rho_{la} = \bar{\alpha_{\cdot l}} z_{ola}/\eta_{oa},$$

Here, δ_{an} is the Kronecker delta, equal to 1 when $n = a$, and equal to 0 when $n \neq a$.

For simplicity, we will start by assuming that all of the $\mathbf{x}_{oi}s$ have common variation matrix $\mathbf{T}$. If the assumptions of multivariate normality are valid for $\mathbf{x}_i$, $\mathbf{z}_{ji}$ and $\boldsymbol{\epsilon}_j$, then based on Approximation 3 in Section 2.6, we can estimate the mean and variance of $\mathbf{Y}_j$ and assume it's normality, which gives us the following information about the components of our model:

$$\boldsymbol{\epsilon}_j \sim \text{MVN}^{D-1}(\mathbf{0}, -\frac{1}{2}\mathbf{V}^T\mathbf{G}\mathbf{T}_\epsilon\mathbf{G}\mathbf{V}) \tag{3.9}$$

$$\begin{aligned}
\mathbf{Z}_{ji} &\sim \text{MVN}^{D-1}(\boldsymbol{\mu}_i, -\frac{1}{2}\mathbf{V}^T\mathbf{G}\mathbf{T}\mathbf{G}\mathbf{V}), \qquad i = 1, ..., I \\
\mathbf{Y}_j &\sim \text{MVN}^{D-1}(\boldsymbol{\eta}_j, -\frac{1}{2}\mathbf{V}^T\mathbf{G}\mathbf{T}_\epsilon\mathbf{G}\mathbf{V}) \\
\mathbf{Y}_j|\mathbf{Z}_{j1}, ..., \mathbf{Z}_{jI} &\sim \text{MVN}^{D-1}(\boldsymbol{\eta}^*_j, -\frac{1}{2}\mathbf{V}^T\mathbf{G}(\boldsymbol{\Theta} + \mathbf{T}_\epsilon)\mathbf{G}\mathbf{V}) \\
\mathbf{X}_i &\sim \text{MVN}^{D-1}(\boldsymbol{\mu}_i, -\frac{1}{2}\mathbf{V}^T\mathbf{G}\mathbf{T}\mathbf{G}\mathbf{V}), \qquad , i = 1, ..., I
\end{aligned} \tag{3.10}$$

where $\boldsymbol{\mu}_i$ is the population mean of all ilr transformed FA signatures of prey species i, $\boldsymbol{\eta}^*_{oj}$ depends on the unobserved random effects as $\boldsymbol{\eta}^*_{oj} = \sum_{i=1}^{I} \alpha_i z_{oji}$, $\boldsymbol{\eta}^*_j$ is the ilr transformation of $\boldsymbol{\eta}^*_{oj}$, and $\boldsymbol{\eta}_j$ is the marginal mean of $\mathbf{Y}_j$. Note, although it is not pursued here, skew-normal distributions, described in Section 2.4, could be used in place of multivariate normal distributions. The α-transformation, defined in Tsagris et al. (2011), could also be used, which is more general than the ilr transformation. Recall that, $-\frac{1}{2}\mathbf{G}\boldsymbol{\Theta}\mathbf{G}$ is the variance-covariance matrix of the clr transformed composition with variation matrix $\boldsymbol{\Theta}$, and so $-\frac{1}{2}\mathbf{V}^T\mathbf{G}\boldsymbol{\Theta}\mathbf{G}\mathbf{V}$ is the variance-covariance matrix of the ilr transformed composition with variation matrix $\boldsymbol{\Theta}$ (these relationships are described in Equation 2.7).

Because multivariate normal distributions are assumed for transformed prey FA signatures, we need to make sure that this is a valid assumption. First, normal probability plots were examined for each prey species, for all FAs. While some FAs appeared to be following a normal distribution, many were not. In order to correct for this, the *Winsorize* function in the package *DescTools* in R (Signorell (2019)) can be used for each individual FA for each species. This function replaces values above and below an upper and lower bound respectively, with those bound values. That is:

$$w(x) = \begin{cases} l & \text{if } x < l \\ x & \text{if } l < x < u \\ u & \text{if } x > u \end{cases} \tag{3.11}$$

After those values are replaced, the compositions are then divided by the sum of the vector, to ensure they still sum to 1. For our data, we chose to use the upper and lower bounds as the fences of a boxplot. That is, the lower bound is $Q_1 - 1.5IQR$ and the upper bound is $Q_3 + 1.5IQR$, where Q_1 and Q_3 are the first and third quartiles respectively, and $IQR = Q_3 - Q_1$ is the interquartile range.

The normal probability plots before and after winsorizing for all species were compared. Two of such comparisons are shown in Figures 3.1 and 3.2. Pollock shows that many FAs are roughly following the normal distribution without winsorizing, but once the FAs are winsorized, there is an improvement, particularly seen with 20:3n-6 and 22:4n-6.For these cases where extreme outliers were present, winsorizing brought these outliers closer into the bulk of the data which increased the normality of the FAs. For Squid, without winsorizing, not many FAs are following the diagonal in the normal probability plots, as many appear to be quite curved and have extreme outliers. However, after winsorizing, some of these curves appear to be straighter and the outliers less severe, as seen with 20:4n-3, 21:5n-3 and 22:4n-3.

The function *shapiro.test* in the package *stats* in R performs the Shapiro-Wilk test of normality, described in Royston (1995), to test for normality. This function was used to obtain p-values for every FA, for every prey species, where p-values larger than the significance levels are indicative of near normality. The proportions of such p-values over several common significance levels were explored both before winsorizing and after. These are shown in Table 3.1.

	Significance		
Winsorized	0.01	0.05	0.1
No	0.7946	0.5982	0.5417
Yes	0.8661	0.7143	0.6161

Table 3.1: Proportions of Shapiro-Wilk test p-values that are above the significance levels listed, for both the winsorized and non-winsorized preybases.

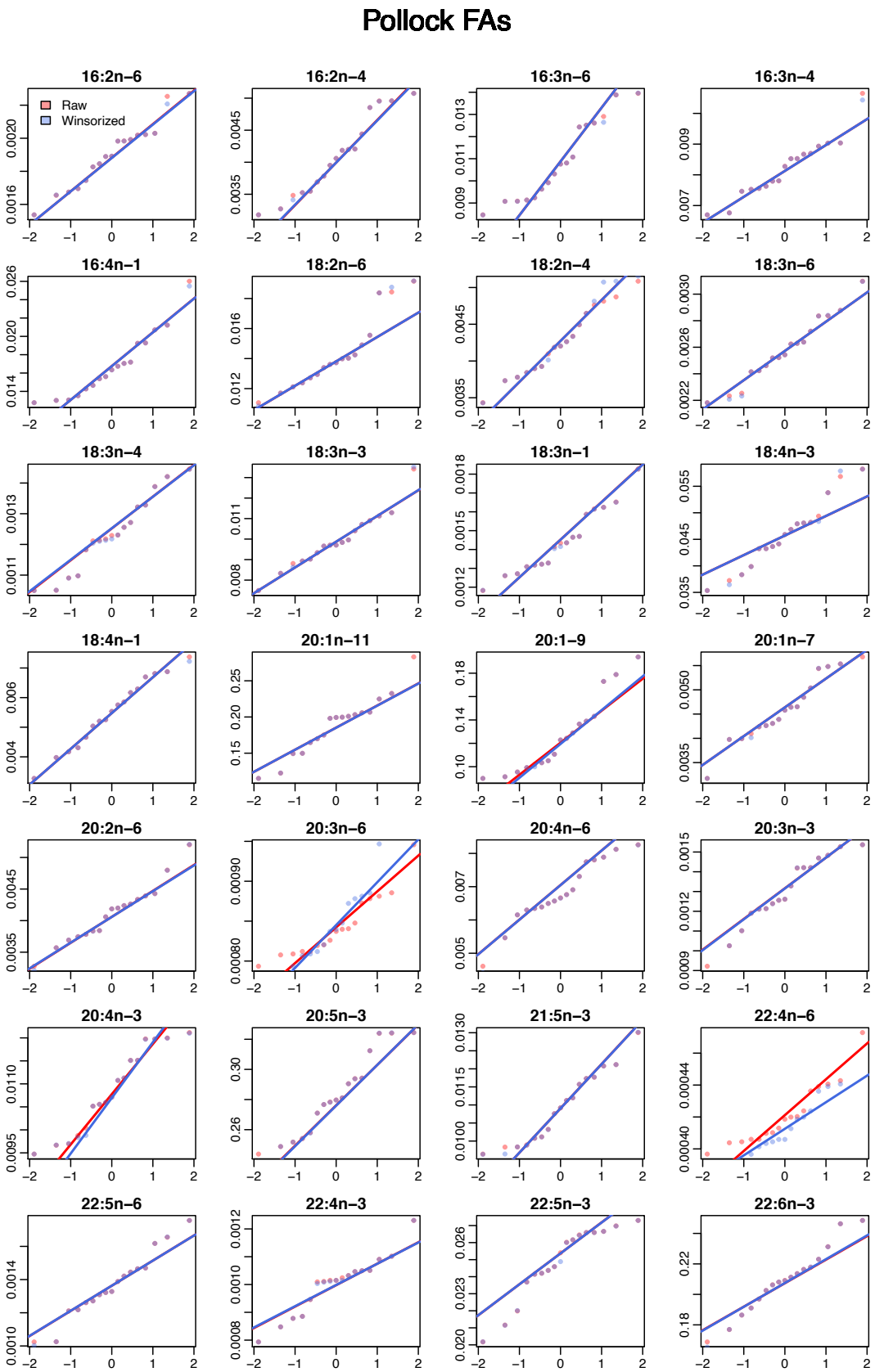

Figure 3.1: Normal probability plots of each FA for Pollock both with (blue) and without (red) winsorizing.

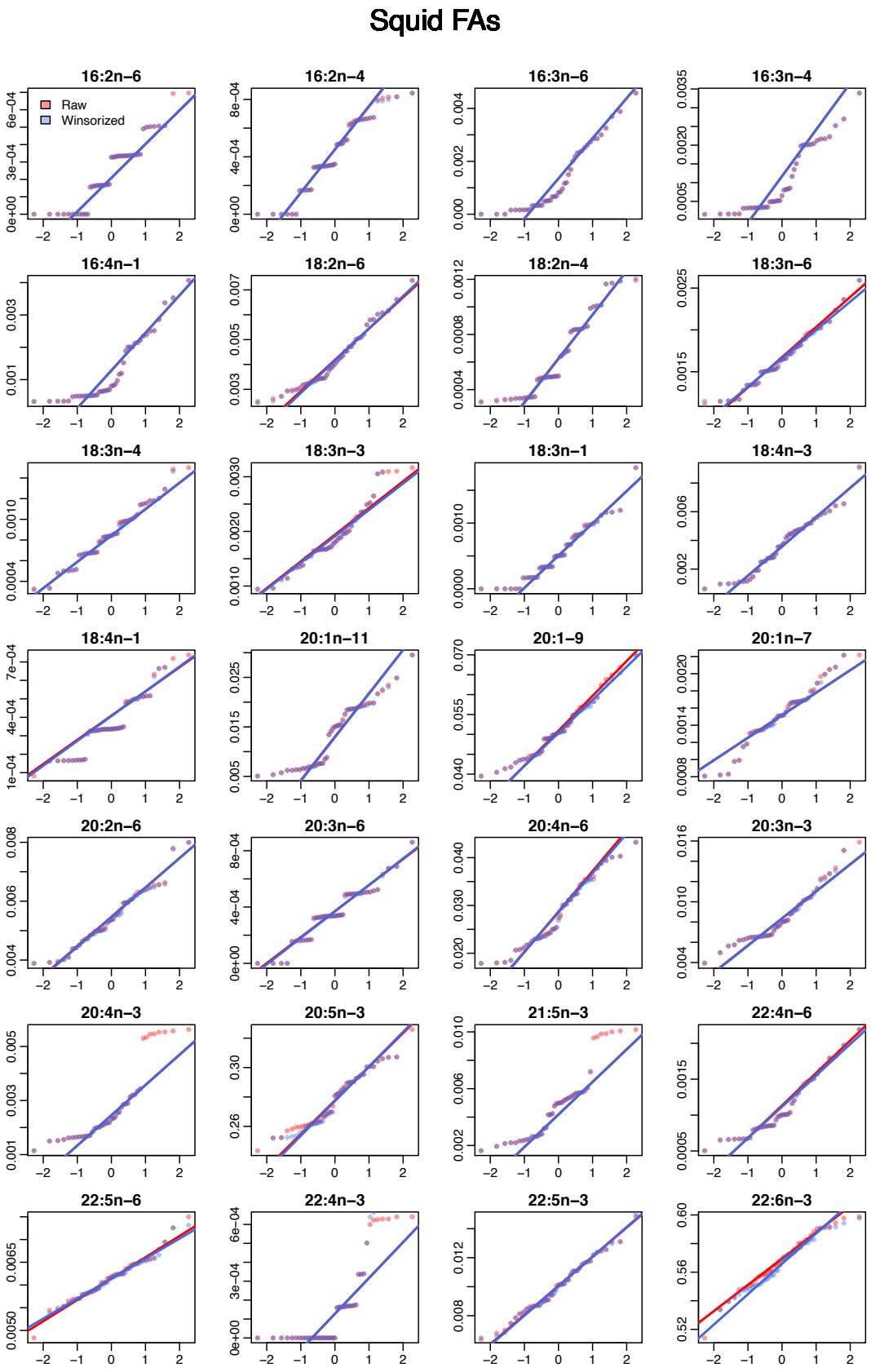

Figure 3.2: Normal probability plots of each FA for Squid both with (blue) and without (red) winsorizing.

We can see from Table 3.1 that winsorizing improves the normality of the FAs for the prey species. Therefore, for non-parametric simulations, and real life data, we will be explore winsorizing the preybase, and by comparing to non-winsorized preybase estimates, see if this affects the accuracy of diet estimation.

We can make the assumption that the predators are conditionally independent, given the random effects, since they will be sampling their own prey species. Similarly we can assume both the observed and unobserved prey are independent of each other. Since the predator is consuming the random effects, and has not consumed the observed prey species, we can assume independence between the predator and observed prey. Finally, we can assume that the random perturbations are independent of each other. Therefore we have:

- $\mathbf{z}_{ji}, i = 1, ..., I$ are independent
- $\mathbf{z}_{ji}, i = 1, ..., I$ and $\boldsymbol{\alpha}_j$ are independent
- $\mathbf{y}_j|\boldsymbol{\alpha}_j, \mathbf{z}_{ji}, j = 1, ..., n.pred, i = 1, ..., I$, are independent
- $\mathbf{x}_1, ..., \mathbf{x}_I$ are independent
- $\mathbf{y}_j$ and $\mathbf{x}_i$ are independent
- $\mathbf{z}_{ji}$ and $\mathbf{x}_i$ are independent
- $\boldsymbol{\epsilon}_j$, $j = 1, .., n.pred$ are independent.

We can find the joint likelihood of $\mathbf{Y}_j$ and $\mathbf{Z}_j$ by considering the density of $\mathbf{Y}_j|\mathbf{Z}_j$ and multiplying by the density of $\mathbf{Z}_j$. So, we need to determine the distribution of $\mathbf{Y}_j|\mathbf{Z}_j$. $\mathbf{Y}_j$ depends on $\mathbf{Z}_j = (\mathbf{Z}_{j1}, \mathbf{Z}_{j2}, ..., \mathbf{Z}_{jI})$, $\boldsymbol{\alpha}_j = (\alpha_{j1}, \alpha_{j2}, ..., \alpha_{jI})$ and $\boldsymbol{\epsilon}_j$. Since we are conditioning on $\mathbf{Z}_j$, these are fixed, as are $\boldsymbol{\alpha}_j$. Therefore, the only randomness involved in $\mathbf{Y}_j|\mathbf{Z}_j$ comes from $\boldsymbol{\epsilon}_j$. So, we have,

$$\begin{aligned} \mathbf{Y}_{oj}|\mathbf{Z}_{oj} &= \boldsymbol{\eta}^*_{oj} \circ \boldsymbol{\epsilon}_{oj} \\ ilr(\mathbf{Y}_{oj}|\mathbf{Z}_{oji}) &= ilr(\boldsymbol{\eta}^*_{oj} \circ \boldsymbol{\epsilon}_{oj}) \\ \mathbf{Y}_j|\mathbf{Z}_j &= ilr(\boldsymbol{\eta}^*_{oj}) + ilr(\boldsymbol{\epsilon}_{oj}) \\ &= \boldsymbol{\eta}^*_j + \boldsymbol{\epsilon}_j \end{aligned} \tag{3.12}$$

where $\boldsymbol{\eta}^*_{oj} = \sum_{i=1}^{I} \alpha_{ji}\mathbf{z}_{oji}$. Since $\text{ilr}(\boldsymbol{\eta}^*_{oj}) = \boldsymbol{\eta}^*_j$ depends only on fixed $\boldsymbol{\alpha}$ and fixed $\mathbf{z}_{oji}$, it is a constant, and $\boldsymbol{\epsilon}_j$ is multivariate Normal, as described in Equation 3.9. Therefore, $\mathbf{Y}_j|\mathbf{Z}_j$ is multivariate Normal, with mean $\boldsymbol{\eta}^*_j + E(\boldsymbol{\epsilon}_j) = \boldsymbol{\eta}^*_j + \mathbf{0} = \boldsymbol{\eta}^*_j$. Since $\mathbf{Z}_j$ are given, the variance-covariance matrix only depends on the variation matrix of the random error, $\mathbf{T}_\epsilon$, so the variance-covariance matrix of $\mathbf{Y}_j|\mathbf{Z}_j$ is $\boldsymbol{\Sigma}_\epsilon = -\frac{1}{2}\mathbf{V}^T\mathbf{G}\mathbf{T}_\epsilon\mathbf{G}\mathbf{V}$. Thus,

$$\mathbf{Y}_j|\mathbf{Z}_j \sim \text{MVN}^{D-1}(\boldsymbol{\eta}^*_j, \boldsymbol{\Sigma}_\epsilon) \tag{3.13}$$

We now need to find the marginal log likelihood (Equation 5.2). If we let $\mathbf{X} = (\mathbf{X}_1, \mathbf{X}_2, ..., \mathbf{X}_I)$, then from the assumptions above, we can write out the density function for the predator FA signatures, conditioning on the random effects, diet proportions, variation matrix of the prey, variation matrix of the random error, and observed prey FA signatures, $f(\mathbf{Y}_j|\mathbf{Z}_j, \boldsymbol{\alpha}_j, \mathbf{T}, \mathbf{T}_\epsilon, \mathbf{X})$.

$$\begin{aligned}
\mathcal{L} &= f(\mathbf{Y}_j|\mathbf{Z}_j, \boldsymbol{\alpha}_j, \mathbf{T}, \mathbf{T}_\epsilon, \mathbf{X}) f(\mathbf{Z}_j) f(\mathbf{X}) \\
&= \frac{1}{(2\pi)^{\frac{D-1}{2}}|-\frac{1}{2}\mathbf{V}^T\mathbf{G}\mathbf{T}_\epsilon\mathbf{G}\mathbf{V}|^{1/2}} \exp\left\{-\frac{1}{2}(\mathbf{Y}_j - \boldsymbol{\eta}^*_j)^T\left[-\frac{1}{2}\mathbf{V}^T\mathbf{G}\mathbf{T}_\epsilon\mathbf{G}\mathbf{V}\right]^{-1}(\mathbf{Y}_j - \boldsymbol{\eta}^*_j)\right\} \times \\
&\prod_{i=1}^{I} \frac{1}{(2\pi)^{\frac{D-1}{2}}|-\frac{1}{2}V^T\mathbf{G}\mathbf{T}\mathbf{G}\mathbf{V}|^{1/2}} \exp\left\{-\frac{1}{2}(\mathbf{Z}_{ji} - \boldsymbol{\mu}_i)^T\left[-\frac{1}{2}V^T\mathbf{G}\mathbf{T}\mathbf{G}\mathbf{V}\right]^{-1}(\mathbf{Z}_{ji} - \boldsymbol{\mu}_i)\right\} \times \\
&\prod_{i=1}^{I} \frac{1}{(2\pi)^{\frac{D-1}{2}}|-\frac{1}{2}\mathbf{V}^T\mathbf{G}\mathbf{T}\mathbf{G}\mathbf{V}|^{1/2}} \exp\left\{-\frac{1}{2}(\mathbf{X}_i - \boldsymbol{\mu}_i)^T\left[-\frac{1}{2}\mathbf{V}^T\mathbf{G}\mathbf{T}\mathbf{G}\mathbf{V}\right]^{-1}(\mathbf{X}_i - \boldsymbol{\mu}_i)\right\}
\end{aligned}$$

For the optimization, many parameters are being estimated simultaneously, particularly when many prey species are being included in the model. Therefore, to minimize the number of parameters in the optimization, $\boldsymbol{\mu}_i$ is estimated using the empirical mean of the observed FA signatures of prey species i, say $\bar{\mathbf{x}}_i$, and $\boldsymbol{\Sigma} = -\frac{1}{2}\mathbf{V}^T\mathbf{G}\mathbf{T}\mathbf{G}\mathbf{V}$ is estimated using the pooled empirical variance-covariance matrices of the ilr transformed prey FA signatures from the prey base, $\hat{\boldsymbol{\Sigma}}$. This is found by estimating the variation matrices for each prey species, $\hat{\mathbf{T}}_i$, converting each to the variance-covariance

matrix of the ilr transformed prey FAs using Equation 2.7 and pooling these estimates. $\mathbf{T}_\epsilon$ is one of the parameters to be estimated during the optimization. Since we have a direct transformation between $\mathbf{T}_\epsilon$ and the variance-covariance matrix of the ilr transformed errors, $\mathbf{\Sigma}_\epsilon$ from Equation 2.7 ($\mathbf{\Sigma}_\epsilon = -\frac{1}{2}\mathbf{V}^T\mathbf{G}\mathbf{T}_\epsilon\mathbf{G}\mathbf{V}$), estimating $\mathbf{\Sigma}_\epsilon$ instead will yield the same results. We will assume that $\mathbf{\Sigma}_\epsilon$ is a diagonal matrix, with $D-1$ values on the diagonal, however estimating all $D-1$ values has proven difficult. Therefore, for the time being, we are splitting the diagonal values into 4 quartiles, and estimating the average of each quartile. This is explained in more detail below. So, the joint likelihood is

$$\begin{aligned}
\mathcal{L} &= f(\mathbf{Y}_j|\mathbf{Z}_j, \boldsymbol{\alpha}_j, \mathbf{T}, \mathbf{T}_\epsilon, \mathbf{X}) f(\mathbf{Z}_j) f(\mathbf{X}) \\
&= \frac{1}{(2\pi)^{\frac{D-1}{2}}|\mathbf{\Sigma}_\epsilon|^{1/2}} \exp\left\{-\frac{1}{2}(\mathbf{Y}_j - \boldsymbol{\eta}_j^*)^T \left[\mathbf{\Sigma}_\epsilon\right]^{-1} (\mathbf{Y}_j - \boldsymbol{\eta}_j^*)\right\} \times \\
&\prod_{i=1}^{I} \frac{1}{(2\pi)^{\frac{D-1}{2}}|\hat{\mathbf{\Sigma}}|^{1/2}} \exp\left\{-\frac{1}{2}(\mathbf{Z}_{ji} - \bar{\mathbf{x}}_i)^T \hat{\mathbf{\Sigma}}^{-1} (\mathbf{Z}_{ji} - \bar{\mathbf{x}}_i)\right\} \times \\
&\prod_{i=1}^{I} \frac{1}{(2\pi)^{\frac{D-1}{2}}|\hat{\mathbf{\Sigma}}|^{1/2}} \exp\left\{-\frac{1}{2}(\mathbf{X}_i - \bar{\mathbf{x}}_i)^T \hat{\mathbf{\Sigma}}^{-1} (\mathbf{X}_i - \bar{\mathbf{x}}_i)\right\}
\end{aligned}$$

The last row of the likelihood above is the density of the prey FA signatures, $f(\mathbf{X})$. This density does not depend on any of the parameters α_i or $\mathbf{\Sigma}_\epsilon$, nor does it depend on the random effects $\mathbf{Z}_j$. Therefore, it is constant relative to our parameters, and is not needed in the likelihood. So, the likelihood to be optimized is:

$$\begin{aligned}
\mathcal{L} &= f(\mathbf{Y}_j|\mathbf{Z}_j, \boldsymbol{\alpha}_j, \mathbf{T}, \mathbf{T}_\epsilon, \mathbf{X}) f(\mathbf{Z}_j) \\
&= \frac{1}{(2\pi)^{\frac{D-1}{2}}|\mathbf{\Sigma}_\epsilon|^{1/2}} \exp\left\{-\frac{1}{2}(\mathbf{Y}_j - \boldsymbol{\eta}_j^*)^T \mathbf{\Sigma}_\epsilon^{-1} (\mathbf{Y}_j - \boldsymbol{\eta}_j^*)\right\} \times \\
&\prod_{i=1}^{I} \frac{1}{(2\pi)^{\frac{D-1}{2}}|\hat{\mathbf{\Sigma}}|^{1/2}} \exp\left\{-\frac{1}{2}(\mathbf{Z}_{ji} - \bar{\mathbf{x}}_i)^T \hat{\mathbf{\Sigma}}^{-1} (\mathbf{Z}_{ji} - \bar{\mathbf{x}}_i)\right\}
\end{aligned}$$

Note, the likelihood above is for the j^{th} individual predator. Since the predator FAs

are assumed to be independent, we can get the joint likelihood of multiple predators simply by multiplying the likelihoods, as seen below.

$$\begin{aligned}\mathcal{L} &= \prod_{j=1}^{n.pred} f(\mathbf{Y}_j|\mathbf{Z}_j, \boldsymbol{\alpha}_j, \mathbf{T}, \mathbf{T}_\epsilon, \mathbf{X}) f(\mathbf{Z}_j) \\ &= \prod_{j=1}^{n.pred} \left(\frac{1}{(2\pi)^{\frac{D-1}{2}} |\boldsymbol{\Sigma}_\epsilon|^{1/2}} \exp\left\{ -\frac{1}{2} (\mathbf{Y}_j - \boldsymbol{\eta}_j^*)^T \boldsymbol{\Sigma}_\epsilon^{-1} (\mathbf{Y}_j - \boldsymbol{\eta}_j^*) \right\} \times \right. \\ &\left. \prod_{i=1}^{I} \frac{1}{(2\pi)^{\frac{D-1}{2}} |\hat{\boldsymbol{\Sigma}}|^{1/2}} \exp\left\{ -\frac{1}{2} (\mathbf{Z}_{ji} - \bar{\mathbf{x}}_i)^T \hat{\boldsymbol{\Sigma}}^{-1} (\mathbf{Z}_{ji} - \bar{\mathbf{x}}_i) \right\} \right)\end{aligned} \tag{3.14}$$

We now need to integrate out the random effects to obtain the marginal log likelihood which is given in Equation 3.15.

$$\begin{aligned}\mathcal{L} &= \int \cdots \int \prod_{j=1}^{n.pred} f(\mathbf{Y}_j|\mathbf{Z}_j, \boldsymbol{\alpha}_j, \mathbf{T}, \mathbf{T}_\epsilon, \mathbf{X}) f(\mathbf{Z}_j) d\mathbf{Z}_{j1} \cdots d\mathbf{Z}_{jI} \\ &= \int \cdots \int \prod_{j=1}^{n.pred} \left(\frac{1}{(2\pi)^{\frac{D-1}{2}} |\boldsymbol{\Sigma}_\epsilon|^{1/2}} \exp\left\{ -\frac{1}{2} (\mathbf{Y}_j - \boldsymbol{\eta}_j^*)^T \boldsymbol{\Sigma}_\epsilon^{-1} (\mathbf{Y}_j - \boldsymbol{\eta}_j^*) \right\} \times \right. \\ &\left. \prod_{i=1}^{I} \frac{1}{(2\pi)^{\frac{D-1}{2}} |\hat{\boldsymbol{\Sigma}}|^{1/2}} \exp\left\{ -\frac{1}{2} (\mathbf{Z}_{ji} - \bar{\mathbf{X}}_i)^T \hat{\boldsymbol{\Sigma}}^{-1} (\mathbf{Z}_{ji} - \bar{\mathbf{X}}_i) \right\} \right) d\mathbf{Z}_{j1} \cdots d\mathbf{Z}_{jI}\end{aligned} \tag{3.15}$$

The R package "TMB" (Kristensen et al. (2016)) uses the Laplace approximation to approximate the marginal log likelihood (Equation 3.15) which we can then optimize to obtain the diet proportions, $\boldsymbol{\alpha}$.

3.2.1 Computations

Algorithm for Analysing Predator FA Signatures

As we are dealing with compositional data in the prey FA signatures, the diets, and the predator FA signatures, there are many steps and transformations that need to be performed before analysis. In order to clarify the steps taken in this process, a break down of the algorithm for estimating diet is described below. This explains the steps in R, as well as the steps in C++ (required for TMB), as well as the form of the parameters and data required.

In R

- Set and load inputs for the algorithm.
- (Optional) If using non-parametric simulations, or real life data, winsorize the preybase.
- Modify any zeros using the multiplicative replacement method (Definition 2.14) and transform the prey database using the ilr transformation described in Definition 2.11.
- Calculate the empirical mean for each species' transformed FA signatures.
- Calculate the estimated variance-covariance matrix of the ilr transformed FA signatures for each prey species, and pool.
- Using the predator data and the original QFASA method discussed in Section 1.1, obtain diet estimates.
- Modify any zeros in the prey and predator datasets using the multiplicative replacement method (Definition 2.14 and transform the predator data using the ilr transformation in Definition 2.11.
- Set start values for the parameters:
  - alpha ($\boldsymbol{\alpha}$) - Matrix of dimension n.pred $\times$ $(I-1)$. Starting values for this parameter are QFASA diet estimates, where each row represents the diet of an individual predator, with the last proportion dropped.
  - Z ($\mathbf{z}$) - Array of dimension $n.pred \times I \times (D-1)$. Starting values for the random effects are the mean transformed FA signatures of the prey species. Each element is a matrix, where the row represents the mean random effect for the i^{th} prey species, and the columns represent the FAs.
  - sepsilon (quartered diagonal of $\boldsymbol{\Sigma}$) - Vector of length 4. Represents the diagonal of the variance-covariance matrix of $\boldsymbol{\epsilon}$. This is split into 4 quartiles, and the mean of each quarter is estimated. The first value represents the mean of the lower 25% of the diagonal entries, the second value is the

mean of the 25%-50% entries, and so on. Starting values for the shortened diagonal are obtained from the following steps:

1. Consider the observed FA signatures of the predators as $\mathbf{y}_{oj}$. Obtain diet estimates $\hat{\boldsymbol{\alpha}}_j$ for each predator j using QFASA.
2. Using the diet estimates obtained in the first step, generate a pseudo-predator without error by generating a generating a FA signature $\mathbf{x}_{oij}$ for each prey species i, using the multivariate normal distribution (see Section 2.4), taking a linear combination of these FA signatures, similar to parametric pseudo-predators, but stopping before the error perturbation, $\hat{\boldsymbol{\eta}}_{oj}^{-\epsilon} = \sum_{i=1}^{I} \hat{\alpha}_{ij}\mathbf{x}_{oij}$.
3. Using the inverse perturbation function described in Definition 2.6, obtain estimates for each $\boldsymbol{\epsilon}_{oj}$. $(\hat{\boldsymbol{\epsilon}}_{oj} = \hat{\boldsymbol{\eta}}_{oj}^{-\epsilon-1} \circ \mathbf{y}_{oj})$.
4. Repeat steps 2 and 3 above 50 times for each predator.
5. Transform the estimated error vectors, $\hat{\boldsymbol{\epsilon}}_{oj}$s, using the ilr transformation, and estimate the variance-covariance matrix.
6. Take the diagonal of the variance-covariance matrix, and obtain the quartiles. Obtain the mean of the entries in the lower quarter, second quarter, third quarter and upper quarter. An index vector is created that says which quarter each diagonal element belongs to. The variance-covariance matrix using the starting values could be then obtained by plugging in the mean of the first quantile for all diagonal entries that are in the first quarter, etc.

- Set data list to pass into TMB.
  - Y - Matrix of dimension $n.pred \times (D-1)$. This is the ilr transformed predator FA signatures.
  - n - Vector of size I. The i^{th} element represents the sample size of the i^{th} prey species in the prey base.
  - varz - Matrix of dimension $(D-1) \times (D-1)$. It is the pooled variance-covariance matrix of the ilr transformed prey.

  - mu - Matrix of dimension $I \times (D-1)$. The i^{th} row is the mean ilr transformed FA signature of the i^{th} prey species.
  - V - Matrix of dimension $D \times (D-1)$. It is the clr basis matrix which is output from "ilrBase".
  - sind - Vector of length $D-1$. This is the index vector created in step 6 above, that indicates which quarter each diagonal element in the error variance-covariance matrix belongs to.

- Compile the TMB function, discussed in the next section.
- Obtain the TMB objective function to be optimized (remember that this represents the approximate marginal negative log likelihood of Y, found in Equation 3.15) using "MakeADFun", passing in both the parameter starting values and the data.
- Make a function that obtains the sum of each row of alpha. This will be the inequality function passed to "solnp".
- Optimize the function spit out by TMB by using "solnp", passing in the function from TMB, the starting values from TMB, the inequality function that you set the lower bound to a vector of 0s and the upper bound to a vector of 1s. Since the alpha matrix optimizes all the diet proportions except for the last prey species, the sum will be between 0 and 1, and we can find the last proportion by subtracting the sum of all the proportions from 1. Lower bounds for alphas and sepsilon are set to 0, and upper bounds for alphas and sepsilon are set to 1s and INF respectively. Note, "solnp" is used here for the simple use of a linear equality constraint, without having to code a Lagrange Multiplier in the likelihood.

TMB function in C++

- Initialize all parameters and data sets.
- Initialize negative log likelihood (nll) to zero

- Declare a multivariate normal distribution with covariance matrix varz for X and Z (nll_dist).
- Create a matrix with sepsilon on the diagonal to represent covariance matrix of the errors (and thus $\mathbf{Y}|\mathbf{Z}$, see Section 3.2).
- Declare multivariate normal distribution with covariance matrix equal to the diagonal matrix found above, for $\mathbf{Y}|\mathbf{Z}$ (nll_y, Equation 3.13).
- In a for loop from 0 to $n.pred - 1$, index is w:
  - Extract w^{th} Z matrix from array Z.
  - Back transform the Z matrix so that it is on the untransformed scale, Z_o.
  - Multiply the w^{th} row of alpha by the Z_o matrix to obtain eta, the untransformed mean of Y.
  - Modify any zeros in eta, and ilr transform to get ymean.
  - Calculate the w^{th} row of Y minus the transformed ymean ($(\mathbf{y}_w - \boldsymbol{\eta}^*_w)$ in Equation 3.15), and get the negative log likelihood at this value using nll_y ($f(\mathbf{Y}_w|\mathbf{Z}, \boldsymbol{\alpha}, \mathbf{T}, \mathbf{T}_\epsilon, \mathbf{X})$ from Equation 3.15). Add this to nll.
  - Inside a second for loop, from 0 to $I - 1$, index i
    - Calculate the i^{th} row of Z minus the i^{th} row of mu ($(\mathbf{Z}_i - \bar{\mathbf{x}}_i)$ in Equation 3.15), and get the negative log likelihood of this value using nll_dist ($f(\mathbf{Z}_i)$ from Equation 3.15). Add this to nll.
  - End for loop
- End for loop
- Return nll.

From this algorithm, the code in R and C++ can be written. Replacing zeros and proper transformation must be done in order for the analysis to run and obtain accurate estimates. See Appendix A.1 for full code.

Chapter 4

Simulations

In this chapter, simulation studies are carried out and discussed in order to assess the methods proposed in Chapter 3. The goal of a simulation study is to generate pseudo data based on known parameters that resemble real life data. In this way, estimates can be compared to the true known parameters of the pseudo data to evaluate how well methods perform in a variety of cases. For our methods, we needed to create pseudo-predators that simulate predator FA signatures with known diet proportions. We achieved this in two ways: parametrically, and non-parametrically (see Section 4.2). Then, using a real-life preybase containing sampled FA signatures of potential prey species, ML and QFASA diet estimates were obtained which were then compared to the true diet of that pseudo-predator. The preybase used for the simulations performed in this thesis is described below, followed by detailed descriptions of the simulations performed.

4.1 Preybase

The preybase used for the simulations is the spring Scotian Shelf preybase discussed in Budge et al. (2002), which has been frequently used for FA analysis. Most species were sampled during random bottom-trawl surveys stratified by zones on the Scotian Shelf (Northwest Atlantic Fisheries Organization subareas 4V, 4W, and 4X) and on George's bank (subarea 5Z) in the spring, summer or fall of 1993, 1994, 1995, 1996, or in the southern Gulf of St. Lawrence (subarea 4T) in 1999. Some of the invertebrate species were collected from research trips or commercial fisheries as well. Species were then frozen and stored at -20°C in sealed plastic bags until analysis was performed.

The specimens were thawed and the length and weight were measured. Then, each individual was homogenized in a blender or food processor. The modified version of the method in Folch et al. (1957) was used to extract the lipids from the samples.

Species	Scientific Name	n	avg. FC
American plaice	*Hippoglossoides platessoides*	134	2.33
Atlantic butterfish	*Peprilus triacanthus*	75	10.83
Atlantic cod	*Gadus morhua*	109	2.46
Atlantic herring	*Clupea harengus*	229	6.44
Atlantic mackerel	*Scomber scombrus*	32	5.14
Capelin	*Mallotus villosus*	162	5.21
Longhorn sculpin	*Clupea harengus*	45	2.06
Northern sandlance	*Ammodytes dubius*	148	5.26
*Northern shortfin squid	*Illex illecebrosus*	35	3.01
Pollock	*Pollachius virens*	53	2.44
Redfish	*Sebastes sp.*	54	7.10
Sea raven	*Hemitripterus americanus*	71	1.97
Silver hake	*Merluccius bilinearis*	58	1.60
Smooth skate	*Malacoraja senta*	33	2.53
Snake blenny	*Lumpenus lumpretaeformis*	18	2.43
Thorny skate	*Amblyraja radiata*	83	2.59
White hake	*Urophycis tenuis*	80	1.29
Winter flounder	*Pseudopleuronectes americanus*	50	1.95
Winter skate	*Leucoraja ocellata*	40	1.47
Witch flounder	*Glyptocephalus cynoglossus*	24	1.91
Yellowtail flounder	*Limanda ferruginea*	156	2.21

Table 4.1: Species, sample sizes, and average fat content (%) included in the prey database used in simulations. Asterisk (*) identifies the invertebrate species.

Tissue samples weighing 1.5g were extracted using 30mL of 2:1 chloroform-methanol and then was washed, filtered through anhydrous sodium sulphate, evaporated under nitrogen, and vacuum sonicated to obtain total lipid weight. Following the methods described in Iverson et al. (1997), fatty acid methyl esters (FAME) were prepared and analyzed in duplicate.

This preybase includes 21 different prey species, all of which are listed in Table 4.1 along with the sample size and average lipid, or fat content, measured as percent wet weight. For each prey, 67 FA proportions were measured. The dietary subset of FAs was used (29 FAs), which includes only those which are from consumed prey (those which are not biosynthesized). The list of all FAs, as well as which FAs are included in the dietary subset, are shown in Table 4.2.

FA	Dietary	FA	Dietary	FA	Dietary
12:0		13:0		Iso14	
14:0		14:1n-9		14:1n-7	
14:1n-5		Iso15		Anti15	
15:0		15:1n-8		15:1n-6	
Iso16		16:0		16:1n-11	
16:1n-9		16:1n-7		7Me16:0	
16:1n-5		16:2n-6	X	Iso17	
16:2n-4	X	16:3n-6	X	17:0	
16:3n-4	X	17:1		16:3n-1	
16:4n-3	X	16:4n-1	X	18:0	
18:1n-13		18:1n-11		18:1n-9	
18:1n-7		18:1n-5		18:2d511	
18:2n-7		18:2n-6	X	18:2n-4	X
18:3n-6	X	18:3n-4	X	18:3n-3	X
18:3n-1	X	18:4n-3	X	18:4n-1	X
20:0		20:1n-11	X	20:1n-9	X
20:1n-7	X	20:2n-9		20:2n-6	X
20:3n-6	X	20:4n-6	X	20:3n-3	X
20:4n-3	X	20:5n-3	X	22:1n-11	
22:1n-9		22:1n-7		22:2n-6	
21:5n-3	X	22:4n-6	X	22:5n-6	X
22:4n-3	X	22:5n-3	X	22:6n-3	X
24:1n-9					

Table 4.2: List of FAs measured in the preybase (67), and those included in the dietary subset (29) that is used for analysis.

4.2 Pseudo-Predators

Pseudo-predators are generated FA signatures used in place of real-life predator FA signatures in simulation studies. We generate them based on a known diet that we choose, hereafter called the true diet, so we can see how accurate and precise the diet estimates are from each model. We have two ways of generating these pseudo-predators: parametrically and non-parametrically.

Parametrically

In order to generate pseudo-predators parametrically, we use the multivariate normal distribution, which we have assumed for our real-life prey species after transformation in our model (Equation 3.9). First, using the preybase, we estimate a mean ilr transformed FA signature for each species (could alternatively use median, or another measure of centre), as well as a pooled variance-covariance matrix. Using these empirical estimates, we generate one ilr transformed FA signature from the multivariate normal distribution for each prey species. These transformed FA signatures are back transformed, and a linear combination of these signatures with the "true" diet is performed to obtain the untransformed FA signature of our predator, without error, $\boldsymbol{\eta}_o^{-\epsilon}$.

The error terms, $\boldsymbol{\epsilon}$, are also generated from a multivariate normal distribution (Equation 3.9) with mean $\mathbf{0}$ and variance-covariance a diagonal matrix of 0.001. Since no predator FA signatures were available for this preybase, we could not base this on real life values. Therefore, this value was selected to limit the amount of variability so the FA signatures of the prey do not overlap, making it easier to distinguish between the prey. The errors $\boldsymbol{\epsilon}$ are back transformed, and a perturbation (equivalent to addition on the ilr scale, see Definition 2.6) is taken between them and the FA signature, $\boldsymbol{\eta}_o^{-\epsilon}$, on the original scale. This perturbation yields the FA signature of the pseudo-predator on the untransformed scale.

Non-Parametrically

Non-parametric pseudo-predators create FA signatures generated in such a way that no assumptions are being made on the distribution of the data, so results based on

this method will indicate how the MLE method is performing even if model assumptions are not met. To generate the non-parametric pseudo-predators, a bootstrap sample of prey FA signatures is first collected from each prey species. The number of prey to resample was investigated in Bromaghin (2015), where it was found that using arbitrary sample sizes yields accurate mean FA signatures. Moreover his proposed method achieves this as well as an accurate variation compared to real life data. Because we do not have real life predator FA data for this preybase, we were unable to use his method for our simulations. Therefore, for our simulations, the total sample size of each prey is resampled with replacement and the mean (or median) FA signatures of the bootstrapped samples are calculated for each prey species. A linear combination of these mean prey signatures is then taken with the "true" diet and the signature that results is the FA signature without error, $\boldsymbol{\eta}_o^{-\epsilon}$.

The error term is generated the same way as it was in the parametric simulations, and is then back transformed so that it can be perturbed with the FA signature $\boldsymbol{\eta}_o^{-\epsilon}$. The resulting composition is the FA signature of the pseudo-predator on the untransformed scale.

Simulation Settings

For the simulation study, we wanted to reduce the number of prey species included in the diets in order to speed up computations. Therefore, we looked at the FA signatures of each prey species to select three groups of 4 species with varying levels of similarity. The goal was to have a group of prey species whose FA signatures are very different, a group that is slightly different, and a group that is similar. In order to make this decision, dendrograms were plotted based on Aitchison's (Definition 2.15), chi-squared (Definition 2.19), and KL (Definition 2.18) distances of the mean FA signatures for each prey species. Species that had FA signatures that are similar will be close together on branches, and those having different FA signatures, will be further away. These plots are shown in Figure 4.1. Notice, KL and chi-squared distances yield nearly identical results. Biologists prefer using the KL distance over Aitchison's distance when using QFASA, and chi-squared is a more recently proposed technique that does not require modification of 0s, therefore, those two plots were relied on

more heavily. However, the relationships are similar when using Aitchison's distance as well. Based on these dendrograms, the sets of species selected are shown in Table 4.3.

Computer/Thesis/ErrorSimulations/Dendrocombo.pdf

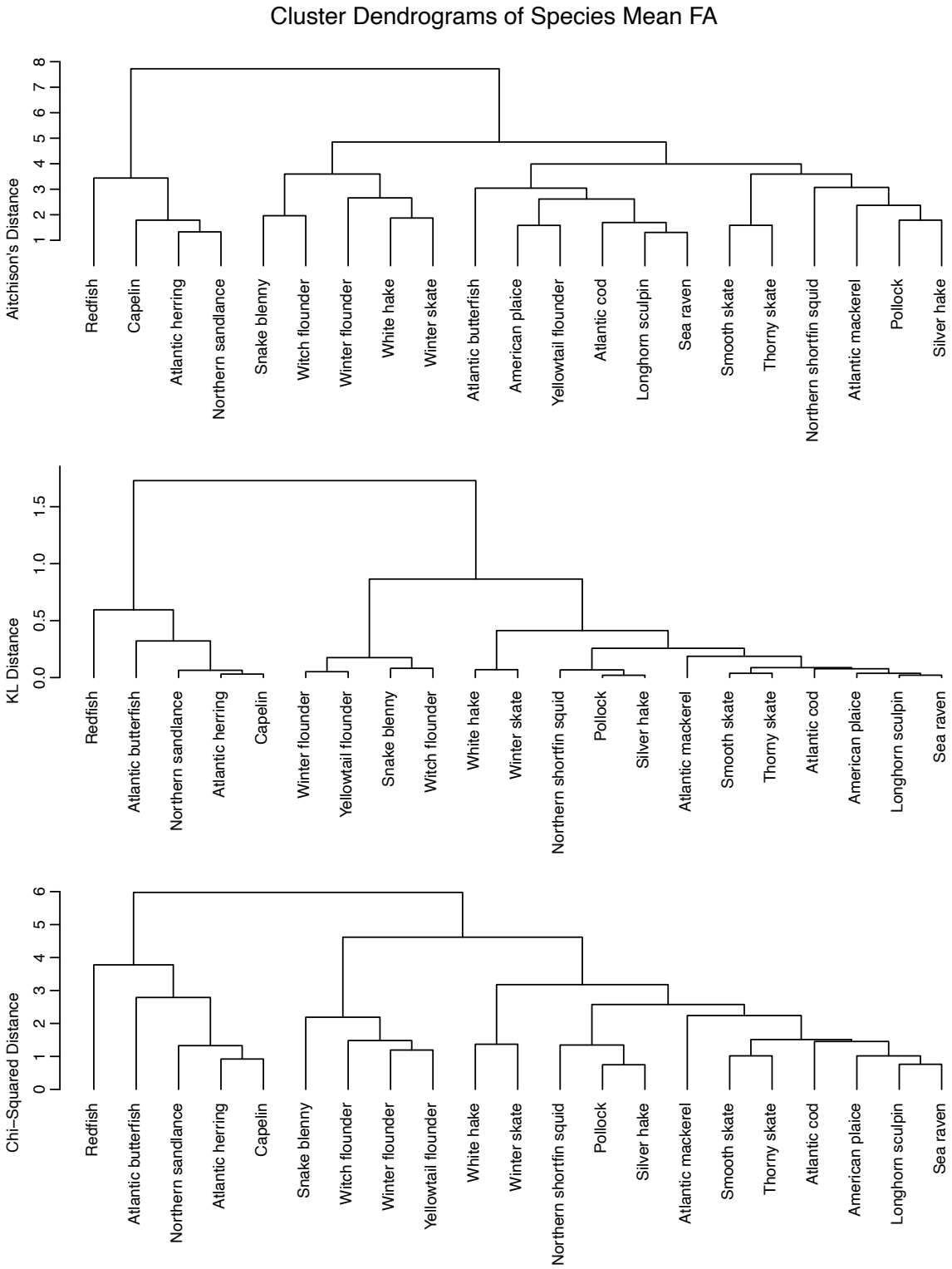

Figure 4.1: Dendrogram using Aitchison's, KL and chi-squared distances of the mean FA signatures for 21 prey species included in the prey data set.

Group	Species 1	Species 2	Species 3	Species 4
Very Different	Capelin	Sea Raven	White Hake	Winter Flounder
Different	Capelin	Pollock	White Hake	Redfish
Similar	Atl. Butterfish	Atl. Herring	Nor. Sandlance	Redfish

Table 4.3: Prey species selected to be included in simulations based on distances between FA signatures.

Groups of 4 species were chosen for each set. Depending on the species of predator for which we are estimating diets, this may or may not be realistic. For grey and harbour seals that we will look at in our real life study, this would be quite a small set, however with the complexity of this model, we wanted to start small and build. Ensuring that this works for smaller sets first, and then adding to our model will save time and potential identifiability issues in the simulations.

For each group of species, regularly-spaced diets across the simplex were used to see how the estimates are performing throughout the whole space. These diets are generated using "*make_diet_grid*" in the package "*qfasar*". Using an increment of $\frac{1}{3}$, 20 diets are generated that are equally spaced by $\frac{1}{3}$ throughout the simplex, which are shown in Table 4.4. These diets ensure that the method is working not only throughout the interior of the simplex, but also on the edges. Between this simulation study and the real life study discussed in Section 6.2, if the ML method is behaving as expected, we could argue that it should function properly in all cases.

Diet	Species 1	Species 2	Species 3	Species 4
1	1	0	0	0
2	0.6667	0.3333	0	0
3	0.6667	0	0.3333	0
4	0.6667	0	0	0.3333
5	0.3333	0.6667	0	0
6	0.3333	0.3333	0.3333	0
7	0.3333	0.3333	0	0.3333
8	0.3333	0	0.6667	0
9	0.3333	0	0.3333	0.3333
10	0.3333	0	0	0.6667
11	0	1	0	0
12	0	0.6667	0.3333	0
13	0	0.6667	0	0.3333
14	0	0.3333	0.6667	0
15	0	0.3333	0.3333	0.3333
16	0	0.3333	0	0.6667
17	0	0	1	0
18	0	0	0.6667	0.3333
19	0	0	0.3333	0.6667
20	0	0	0	1

Table 4.4: Diets spaced equally over the simplex that are included in the simulations.

Each of these diet and species combinations is run with sample size $n = 100$ where pseudo-predators are generated randomly via both methods described, using the dietary subset of FA. QFASA estimates are used as starting values for the diet parameters, which were obtained using calibration coefficients of 1, Aitchison's distance, and the dietary FA subset. For Z, the random effects in Equation 3.6, a matrix of the mean ilr transformed FA signatures for each prey species is used for the starting value.

Simulation Results

Parametric Simulations

With compositional data, Aitchison's distance (Definition 2.15) is often used to measure how far two compositional vectors are from each other. This measure is used as a quantitative way to assess how well the MLE method is performing relative to

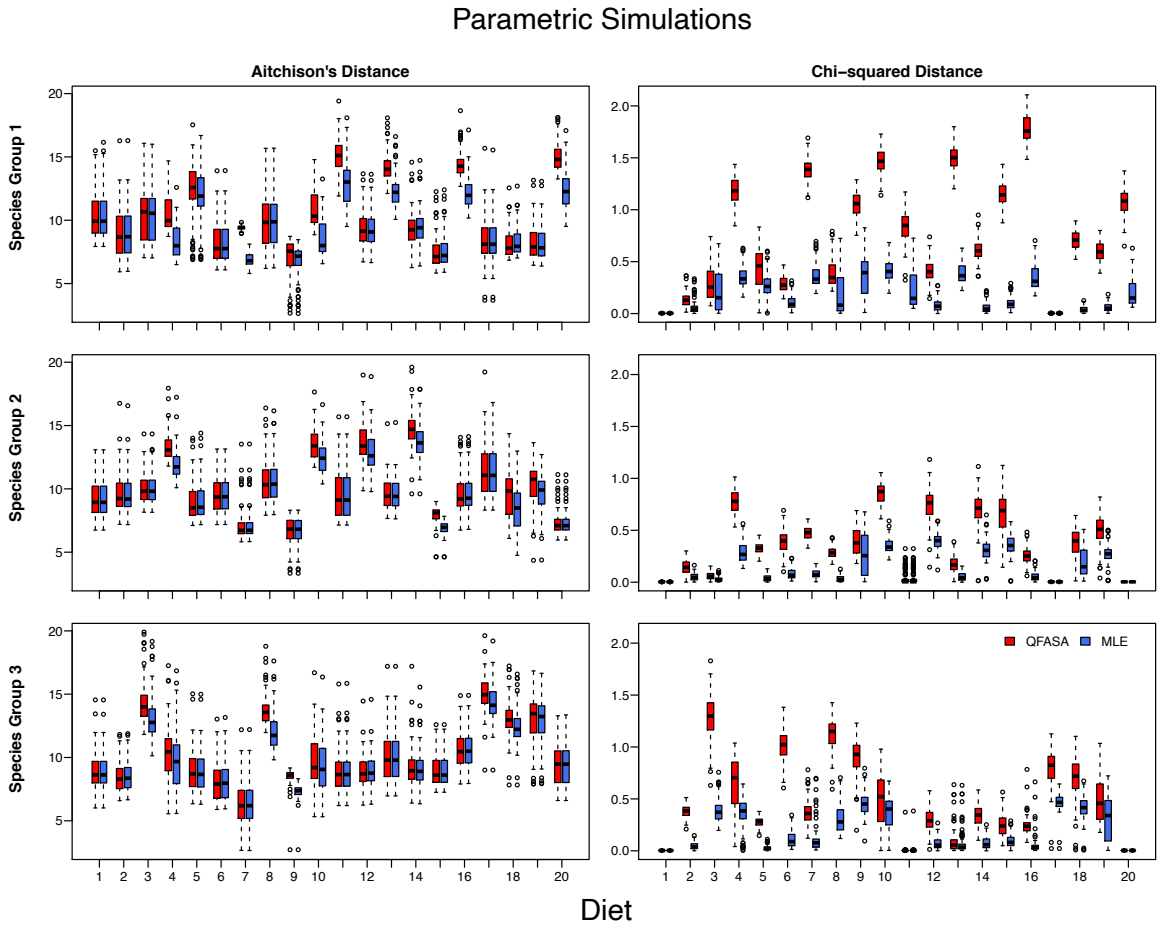

Figure 4.2: Boxplot of Aitchison's and chi-squared distances for 20 diet simulations, with 100 parametric pseudo-predators each, on species groups 1, 2 and 3.

QFASA. In order to use Aitchison's distance, the QFASA and MLE estimates, and the diet vector, were modified using the multiplicative replacement method described in Definition 2.14, as Aitchison's distance cannot be used with 0 elements in the compositions. Then, for each of the 100 predators in all 20 diets, Aitchison's distance was calculated between the estimates (both QFASA and MLE) and the "true" diet. These distances for species group 1 are summarized in Figure 4.2. In this figure, we can see that for all diets, the distances between QFASA estimates and the "true" diet are similar, and in some cases (for example diets 10, 11, 13, 16 and 20) QFASA is much larger than that for the MLE estimates.

The same thing was then done using chi-squared distance (Definition 2.19), proposed

in Stewart (2017). This distance measure is beneficial as it does not require modification of the estimates or true diet. That is, essential zeros, and rounded zeros can be used when calculating the distance between the two compositions. The chi-squared distances for species group 1 are also summarized in Figure 4.2. Using chi-squared distance in Figure 4.2, we can see that nearly all of the QFASA estimates appear to be much further from the true diet than the MLE method.

To explore the reliability of these distances measures, let's look at a simple example. Consider species group 1, diet 6, and the first estimate using both QFASA and MLE. The true diet is (0.333, 0.333, 0.333, 0), and the estimated values are (0.279, 0.292, 0.428, 0.000) and (0.327, 0.306, 0.367, 0.000) using QFASA and MLE, respectively, rounded to 3 decimal places. For QFASA, Aitchison's and chi-squared distances between true and estimated diets are 8.19 and 0.393 respectively and for MLE, they are 8.20 and 0.151. These values are without modification of the zeros in the estimate, as neither method can estimate an essential zero, but always give slightly positive values. However, if we round the estimate to 3 decimal places as displayed above, and replace the zeros using the multiplicative replacement method with $\lambda = 0.00005$, Aitchison's distance for QFASA becomes 0.333 and for MLE, 0.130. Thus, Aitchison's distance relies heavily on the number of decimal points you keep in the composition, and your choice of imputation value, λ. Therefore, chi-squared distances should be considered more reliable when comparing the estimates to the true values, as it does not require any modification of the estimates that could impact the distance measures.

Looking at Table 4.5, we can see that in every combination of species group, and distance measure, QFASA estimates are further away from the true diet on average than those obtained from the MLE method.

We also looked at similar plots for species groups 2 and 3, (all shown within Figure 4.2) and saw similar results; the distances between the true diet and QFASA estimates are generally higher than that with our MLE estimates. What is interesting is that with species group 1, which has FA signatures that are very different from each other, the chi-squared distances between true diet and estimated diet are higher

	Species Group	1		2		3	
Distance	Method	$\bar{d}$	s_d	$\bar{d}$	s_d	$\bar{d}$	s_d
Ait	QFASA	10.4	2.9	10.1	2.6	10.2	2.8
	MLE	9.6	2.4	9.8	2.4	9.9	2.5
Chi	QFASA	0.77	0.53	0.36	0.29	0.49	0.40
	MLE	0.19	0.18	0.15	0.16	0.19	0.20

Table 4.5: Mean ($\bar{d}$) and standard deviation (s_d) of distances (Aitchison's and chi-squared) between the true diet and the estimates from both QFASA and MLE methods, for 3 species groups, with 20 equally spaces diets, and 100 parametric pseudo-predators each.

(ranging between 0 and 2) than with species groups 2 or 3. This is the reverse of what we expected, as different FA signatures should be easier to differentiate between species. This may be a result of the method of selection for the prey species. The dendrogram used only the distance between mean FA signatures. Therefore individual FA signatures could be closer than what was suggested by this plot. Either way, the results are quite favourable, as the MLE method is still performing well relative to QFASA.

When comparing only species groups 2 and 3 in Figure 4.2, we see the expected results, as species group 2, with slightly different FA signatures, have smaller chi-squared distances (ranging between 0 and 1.2) than with species group 3, with similar FA signatures (ranging between 0 and 1.5). When using Aitchison's distance to measure the difference between true and estimated diets, there is not an obvious difference in magnitude from species group to species group. However, we can see that the variability of the estimates seems to decrease as we move from species group 1, to species group 2, and again as we move to species group 3.

To explore the diet estimates one by one, for each true diet, and each species group, boxplots of the 4 prey species' estimates using both MLE and QFASA techniques were plotted with a purple line representing the true diet. This yielded many plots so we will discuss a select few. First, let's consider species group 1. Referring back to Figure 4.2, we can see that using both distance measures, diet 16 had very large distances between true and QFASA estimated diets, but also relatively large distances

between true and MLE estimated diets, whereas diet 17 had very small distances between true and estimated diets using both methods. Boxplots for diet 16 and diet 17 estimates are displayed in Figures 4.3 and 4.4 respectively.

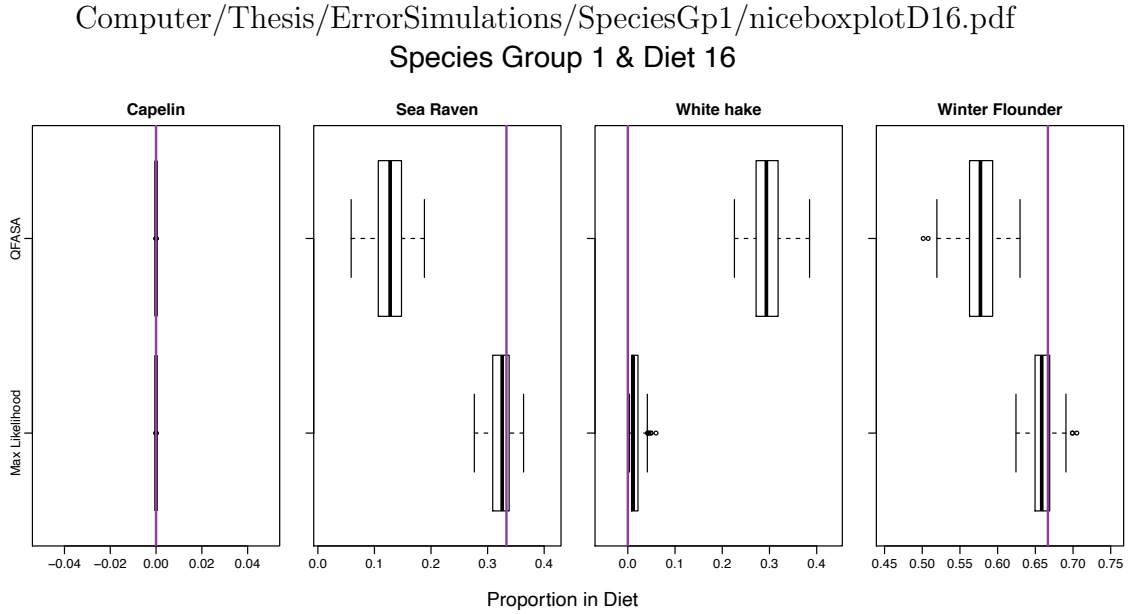

Figure 4.3: Boxplots of estimated diet proportions of 100 parametric pseudo-predators using species group 1 and diet 16 for both QFASA and MLE methods, where true diet is shown in purple.

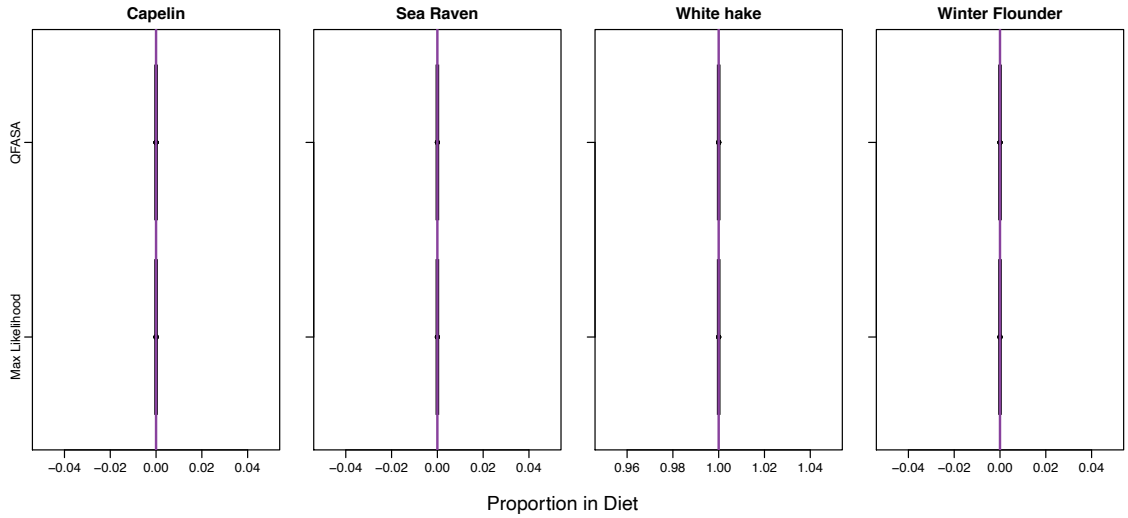

Figure 4.4: Boxplots of estimated diet proportions of 100 parametric pseudo-predators using species group 1 and diet 17 for both QFASA and MLE methods, where true diet is shown in purple.

For diet 16, shown in Figure 4.3, the true diet (0, 0.333, 0, 0.667) is located inward of the edges of the simplex. QFASA had a difficult time estimating all of the proportions other than capelin. It tended to over estimate white hake, and underestimate winter flounder and sea raven. Using the additional measures that "p.QFASA" calculates, the FAs 16:3n-4, 16:2n-6, and 16:3n-6 are contributing the most on average to the Aitchison's distance used to estimate the diets. Together, they account for over 40% of the Aitchison's distance between the predator FA signature, and the linear combination. These FAs have relatively small proportions in the predator signatures, being just slightly larger than the first quartile. Our MLE method does not appear to have this difficulty, as it is not only estimating significantly more accurately than QFASA, but also more precisely, as the variability is also smaller. Those same FAs do not seem to affect our method as much as with QFASA, as the ilr transformation does not inflate their contribution to the FA as the Aitchison's distance does. This shows great promise for our method, especially where this was a case flagged as one of the more poorly estimated diets from the earlier figures.

For diet 17, shown in Figure 4.4, the true diet (0, 0, 1, 0) is on the edge of the simplex. Both MLE methods and QFASA seem to be estimating nearly perfectly for this diet. This is surprising as we expected the difficulty to lie along the edges of the simplex, and not in the centre. Once again, our MLE method is proving to estimate quite successfully when the FA signatures of the prey species are largely different (species group 1).

To compare all the diets for species group 1, a bias table is shown in Table 4.6. Bias of the 100 pseudo-predators is taken by subtracting the estimated diet from the true diet. The bias and standard deviation of the estimates (in parentheses) are shown in the table for each diet in species group 1. Similar to the boxplots, you can see that nearly every diet has smaller mean bias for MLE than for QFASA, sometimes significantly so, and also has a smaller standard deviation in most cases. This reiterates that the MLE is estimating the diets both more accurately and more precisely than QFASA.

Diet	Method	Capelin	Pollock	Redfish	White Hake
1	MLE	0.000(0.000)	0.000(0.000)	0.000(0.000)	0.000(0.000)
	QFASA	0.000(0.00000)	0.000(0.000)	0.000(0.000)	0.000(0.000)
2	MLE	-0.002(0.012)	0.001(0.012)	0.001(0.002)	0.000(0.000)
	QFASA	0.027(0.013)	-0.028(0.014)	0.001(0.002)	0.000(0.000)
3	MLE	-0.006(0.014)	0.009(0.012)	-0.003(0.015)	0.000(0.000)
	QFASA	-0.043(0.0156)	0.012(0.016)	0.0309(018)	0.000(0.000)
4	MLE	-0.009(0.011)	0.006(0.008)	0.011(0.007)	-0.007(0.011)
	QFASA	-0.075(0.019)	0.014(0.019)	0.147(0.032)	-0.086(0.018)
5	MLE	0.000(0.013)	-0.011(0.017)	0.011(0.009)	0.000(0.000)
	QFASA	-0.006(0.016)	-0.021(0.029)	0.027(.022)	0.000 (0.000)
6	MLE	-0.004(0.013)	0.003(0.026)	0.001(0.021)	0.000(0.000)
	QFASA	-0.050(0.015)	-0.004(0.030)	0.054(0.025)	0.000(0.000)
7	MLE	-0.003(0.014)	-0.010(0.026)	0.017(0.012)	-0.005(0.014)
	QFASA	-0.071(0.018)	-0.031(0.035)	0.208(0.033)	-0.106(0.017)
8	MLE	-0.007(0.011)	0.009(0.014)	-0.002(0.014)	0.000(0.000)
	QFASA	-0.074(0.013)	0.009(0.015)	0.065(0.014)	0.000(0.000)
9	MLE	-0.010(0.011)	0.020(0.017)	-0.009(0.019)	-0.001(0.013)
	QFASA	-0.101(0.015)	0.029(0.026)	0.214(0.024)	-0.142(0.016)
10	MLE	-0.009(0.011)	0.008(0.011)	0.014(0.008)	-0.013 (0.0134)
	QFASA	-0.080(0.017)	0.018(0.024)	0.221(0.037)	-0.160(0.027)
11	MLE	0.001(0.004)	-0.011(0.014)	0.010(0.012)	0.000(0.000)
	QFASA	0.001(0.002)	-0.087(0.029)	0.086(0.028)	0.000(0.000)
12	MLE	0.000(0.000)	-0.000(0.023)	0.000(0.023)	0.000(0.000)
	QFASA	0.000(0.000)	-0.100(0.027)	0.100(0.027)	0.000(0.000)
13	MLE	0.000(0.000)	-0.014(0.018)	0.020(0.010)	-0.006(0.014)
	QFASA	0.000(0.000)	-0.173(0.031)	0.244(0.035)	-0.071(0.021)
14	MLE	0.000(0.000)	-0.000(0.018)	0.000(0.018)	0.000(0.000)
	QFASA	0.000(0.000)	-0.135(0.021)	0.135(0.021)	0.000(0.000)
15	MLE	0.000(0.000)	-0.001(0.021)	0.001(0.022)	0.001(0.013)
	QFASA	0.000(0.000)	-0.185(0.023)	0.274(0.026)	-0.089(0.017)
16	MLE	0.000(0.000)	-0.009(0.019)	0.016(0.012)	-0.007(0.017)
	QFASA	0.000(0.000)	-0.205(0.029)	0.295(0.036)	-0.090(0.025)
17	MLE	0.000(0.000)	0.000(0.000)	0.000(0.000)	0.000(0.000)
	QFASA	0.000(0.000)	0.000(0.000)	0.000(0.000)	0.000(0.000)
18	MLE	0.000(0.000)	0.000(0.000)	-0.000(0.012)	0.000(0.012)
	QFASA	0.000(0.000)	0.000(0.000)	0.153(0.017)	-0.153(0.017)
19	MLE	0.000(0.000)	0.000(0.000)	-0.001(0.017)	0.001(0.017)
	QFASA	0.000(0.000)	0.000(0.000)	0.147(0.023)	-0.147(0.023)
20	MLE	0.000(0.000)	0.000(0.000)	0.007(0.008)	-0.007(0.008)
	QFASA	0.000(0.000)	0.000(0.000)	0.137(0.029)	-0.137(0.030)

Table 4.6: Bias and standard deviations (in parentheses) of the estimates for species group 1.

We can do a similar exploration for species groups 2 and 3. First, looking at species group 2 in Figure 4.2, we can see that diet 10 has some of the largest chi-squared and Aitchison's distances between the true and estimated diets, and diet 20 has some of the smallest. The boxplots for diet 10 and diet 20 estimates are shown in Figures 4.5 and 4.6 respectively.

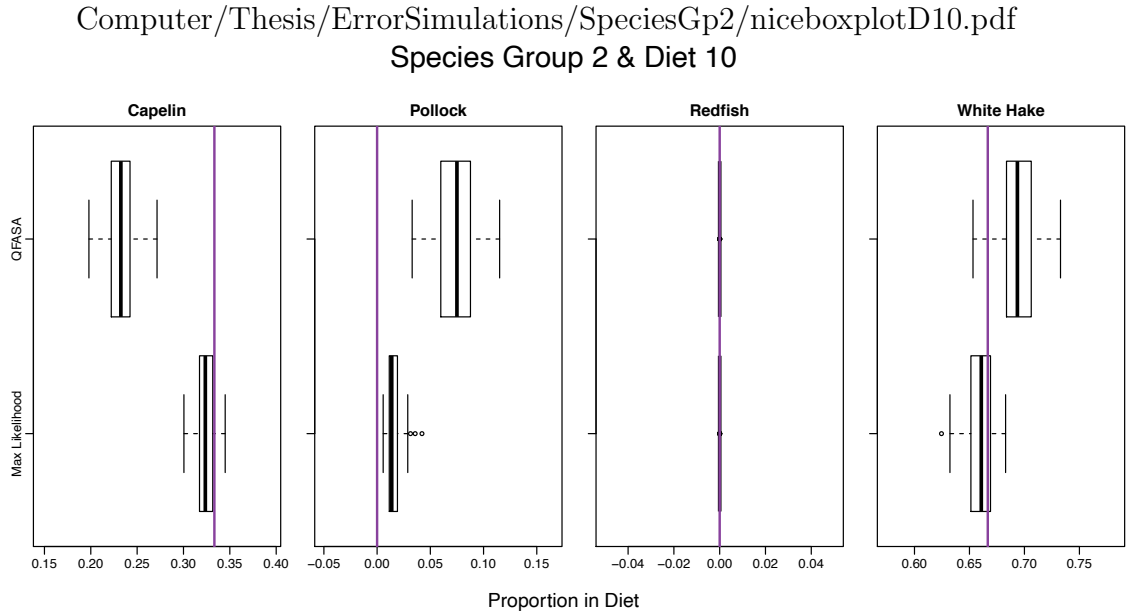

Figure 4.5: Boxplots of estimated diet proportions of 100 parametric pseudo-predators using species group 2 and diet 10 for both QFASA and MLE methods, where true diet is shown in purple.

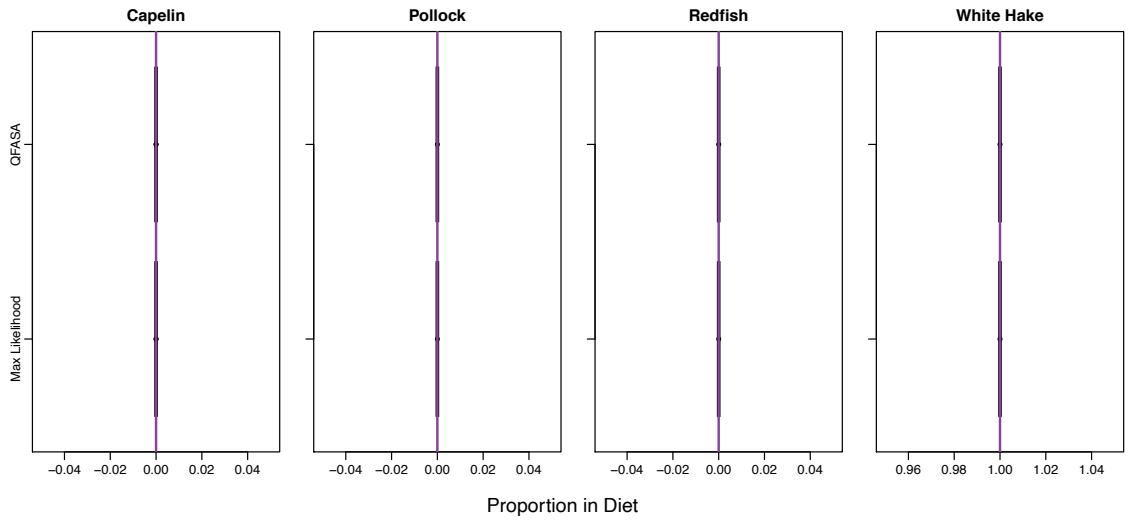

Figure 4.6: Boxplots of estimated diet proportions of 100 parametric pseudo-predators using species group 2 and diet 20 for both QFASA and MLE methods, where true diet is shown in purple.

Similar to species group 1, one of the more poorly estimated diets (depicted in Figure 4.5) has two 0 proportions and a 0.333 and 0.667 proportion, only this time those masses are on different prey species, with a true diet of (0.333, 0, 0, 0.667). QFASA is underestimating capelin by between 5 and 10%, overestimating pollock by between 4 and 12%, and isn't far off, but slightly overestimating on average, for white hake. On average, over 44% of the FA contribution to Aitchison's distance between predator FA signature and the linear combination is due to FA 16:2n-6. The next highest contributing FA explains only 10%. Thus, 16:2n-6 is very influential to the estimates, yet is has the second smallest proportions, on average, in the predator FA signatures and it does not show much variability among the mean FA signature for the four species (SD = 0.0002). Also, out of the top 4 FA contributors to Aitchison's distance, 3 of them have significantly different values for Redfish, which makes Redfish easier to distinguish from the other Prey. These other 3 FAs (16:4n-1, 22:4n-3, 16:3n-4) have standard deviation among the 4 species of 0.0107, 0.0003 and 0.0051 respectively.Therefore, two of the FAs that most contribute to Aitchison's distance have very little variability among the mean proportions. This would make it very difficult

to differentiate between species, thus QFASA could place weight on the wrong species. Since MLE relies on the ilr transformation, which did not put such an extremely large weight on 16:2n-6, once again, our MLE method estimates tend to be within 2% of the true diet, and have a relatively low variability compared with that of QFASA estimates.

One of the better estimates for species group 2, like that of species group 1, is a diet on the edge of the simplex. Shown in Figure 4.6, the true diet of (0, 0, 0, 1) is nearly perfectly estimated with both MLE and QFASA, with little to no variation. Here, the four highest contributing FAs, in decreasing order, are 16:2n-6, 16:3n-6, 16:4n-1, and 16:3n-4. These FAs have standard deviation between the mean FA proportions for the four species, of 0.0002, 0.0065, 0.0107, 0.0051 respectively. These last three have relatively large variability between the species, making it easier for QFASA to distinguish between the 4 species. As well, both methods appear to have an easier time estimating values on the edges of the simplex than those within the simplex, as nearly 100% of the proportion of FAs in the predator is due only to one species FA signature.

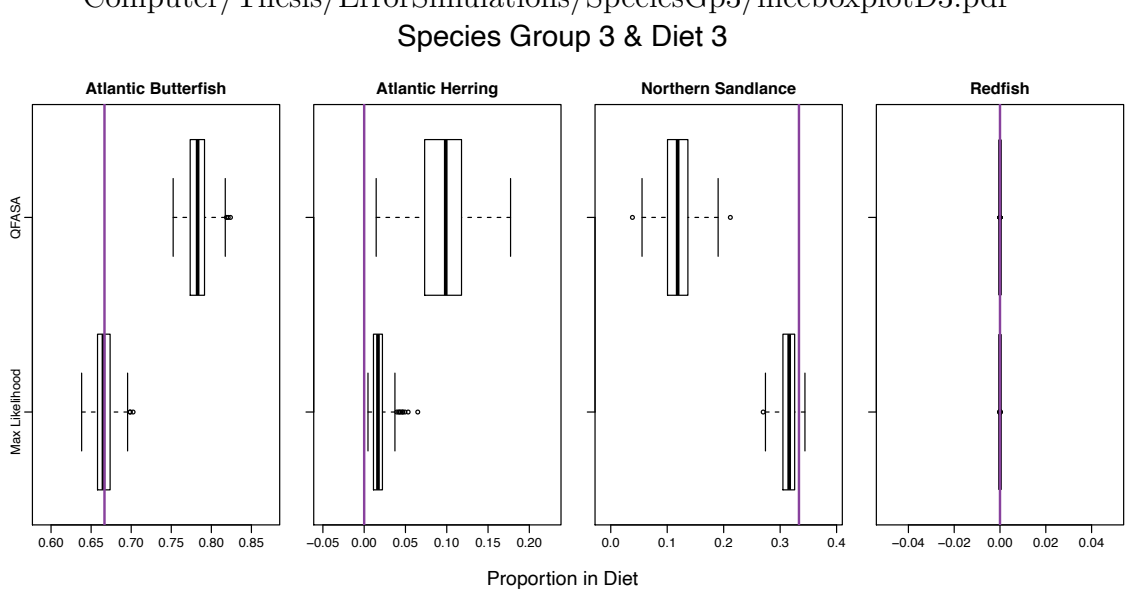

Figure 4.7: Boxplots of estimated diet proportions of 100 parametric pseudo-predators using species group 3 and diet 3 for both QFASA and MLE methods, where true diet is shown in purple.

Looking at species group 3, from the plots in Figure 4.2, diet 3 has some of the largest differences between true and estimated diets, using both chi-squared and Aitchison distances, and diet 7 has some of the smallest. A boxplot of diet 3 is displayed in Figure 4.7, and once again, the true diet here, (0.667, 0, 0.333, 0), is off the edge of one side of the simplex. QFASA is overestimating Atlantic butterfish by 10 to 15%, as well as Atlantic herring by 2 to 18%, the latter being quite a large range. QFASA is underestimating Northern sandlance by between 15-25%. For these estimates, the top 4 FAs contributing the most to Aitchison's distance between predator and linear combination are 16:3n-4, 20:5n-3, 16:4n-1 and 18:2n-6 in decreasing order. With the exception of 20:5n-3, these all have relatively small standard deviations between the mean FA signatures for the four species (0.0039, 0.0366, 0.0076 and 0.0038 respectively) compared to the largest standard deviation (20:1n-9, 0.0925). Therefore, it would be rather difficult to differentiate between the species. However, after removing the mean FA signature for redfish, the standard deviations between the mean FA signatures of the main contributing FAs dramatically decreases (0.0013, 0.0446, 0.0039

and 0.0031 respectively). Therefore, the mean FA signature for redfish is quite distinguishable from that of Atlantic butterfish, Atlantic herring and Northern sandlance, however the other 3 are quite similar. This explains why QFASA is having a difficult time distinguishing the proportions between those three species, but has no difficulty with redfish. The MLE method has a maximum median difference between true and estimated diets of 2%, and relatively small variabilities, with the largest total spread less than 10%. Once again, the ilr transformation used for the ML method, does not place such high weights on the smaller proportions in FA signatures, and thus MLE is estimating, even for species quite similar in FA signatures, more accurately and more precisely than QFASA.

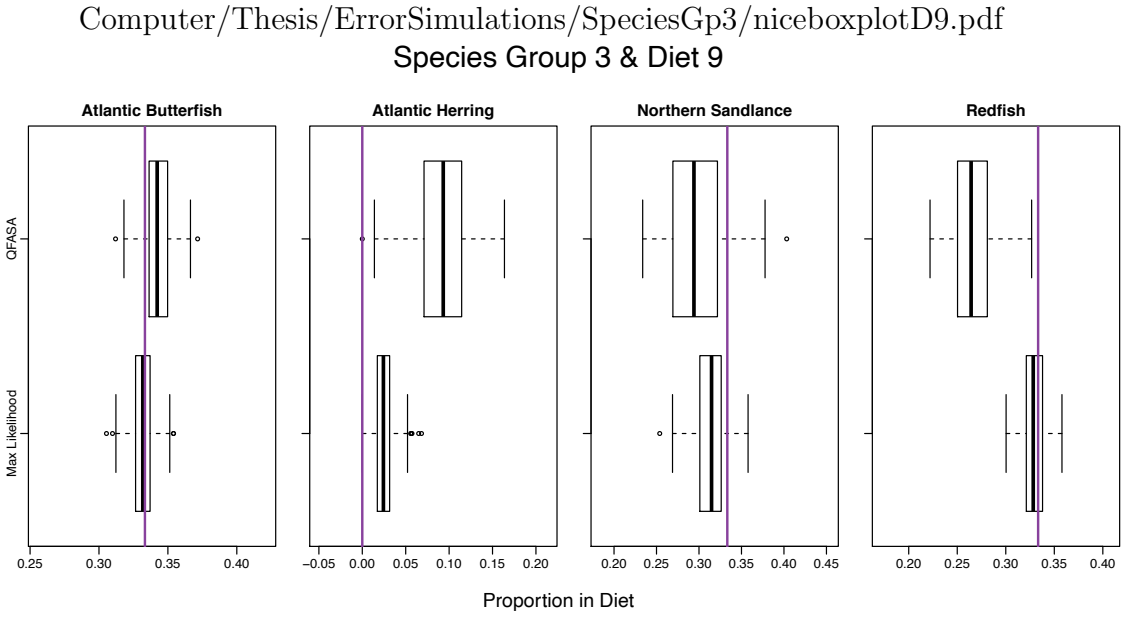

Figure 4.8: Boxplots of estimated diet proportions of 100 parametric pseudo-predators using species group 3 and diet 9 for both QFASA and MLE methods, where true diet is shown in purple.

In Figure 4.8, we can once again see that the MLE method is estimating well in comparison to QFASA. The true diet here is more central in the simplex (0.333, 0, 0.333, 0.333). The median proportion estimate is within 3% of the true diet for all prey species using the MLE method, with relatively small variabilities, the largest having a range of 10%. QFASA is not performing as well here, as it is overestimating Atlantic

herring, as well as Atlantic butterfish on average, and underestimating Northern sandlance and redfish. When looking at the contributions to Aitchison's distance, the four top FAs in decreasing order are 16:4n-1, 16:2n-6, 20:5n-3, nd 16:3n-4. Once again, with the exception of 20:5n-3, they all have relatively small standard deviations, especially the new addition compared to diet 3, 16:2n-6. The standard deviations of the mean FA signatures of the 4 species for those 4 FAs are 0.0076, 0.0004, 0.036 and 0.0039. However, this time, redfish isn't as different for these four FAs, since 16:2n-6 actually has a higher standard deviation (0.0005) after redfish is removed. This explains why QFASA has a difficult time, even with redfish for this diet. All of the estimates also have large variabilities compared to the ML estimates. In summary, for all true diets, and all species groups looked at in this simulation, MLE performed more accurately and precisely than QFASA estimates, when generating parametrically.

Non-Parametric Simulations

For the non-parametric pseudo-predators, the same settings and analyses are performed as with the parametric predators. Since very similar results were seen for all sets, it appears that distance between mean FA signatures of prey has little to no effect on the estimation process. Therefore, only species group 1 is used here.

First, we look at boxplots of the Aitchison's and chi-squared distances between the diet estimates and the true diet. These plots are shown in Figures 4.9 and 4.10 respectively. Note that using Aitchison's distance, both methods have comparable distances to each other, and those using parametric pseudo-predators. However, while using chi-squared distances, the MLE method yields estimates often much further from the true diet than QFASA. To look at this more closely, we once again looked at boxplots of individual diets.

Looking at both figures of distances, we can see that diet 11 appears to have large differences between true diet and estimated diet for both QFASA and MLE methods using both distance measures. Similarly, we can see that diet 7 has relatively small differences. The boxplot depicting diet 11 estimates is displayed in Figure 4.11. For capelin, QFASA is over estimating between 0 and 10%, and MLE is overestimating

Computer/Thesis/ErrorSimulations/Nonparametric/AitchisonDistNP.pdf

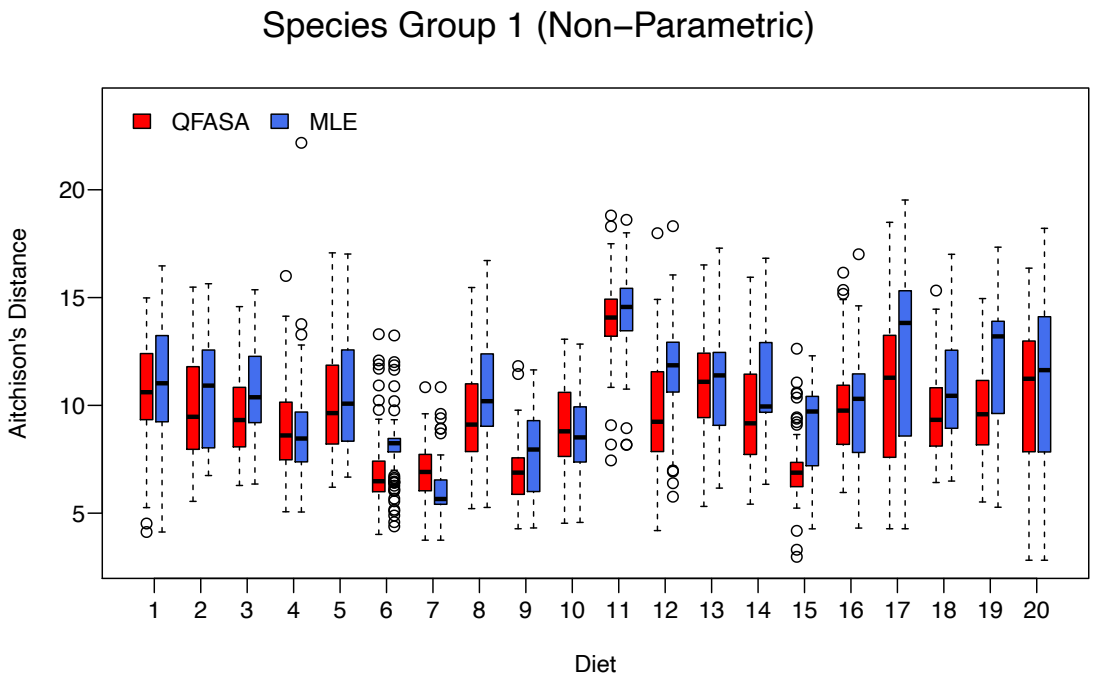

Figure 4.9: Boxplot of Aitchison's distances for 20 diet simulations with 100 non-parametric pseudo-predators each, on species group 1.

Computer/Thesis/ErrorSimulations/Nonparametric/ChiSquaredDistNP.pdf

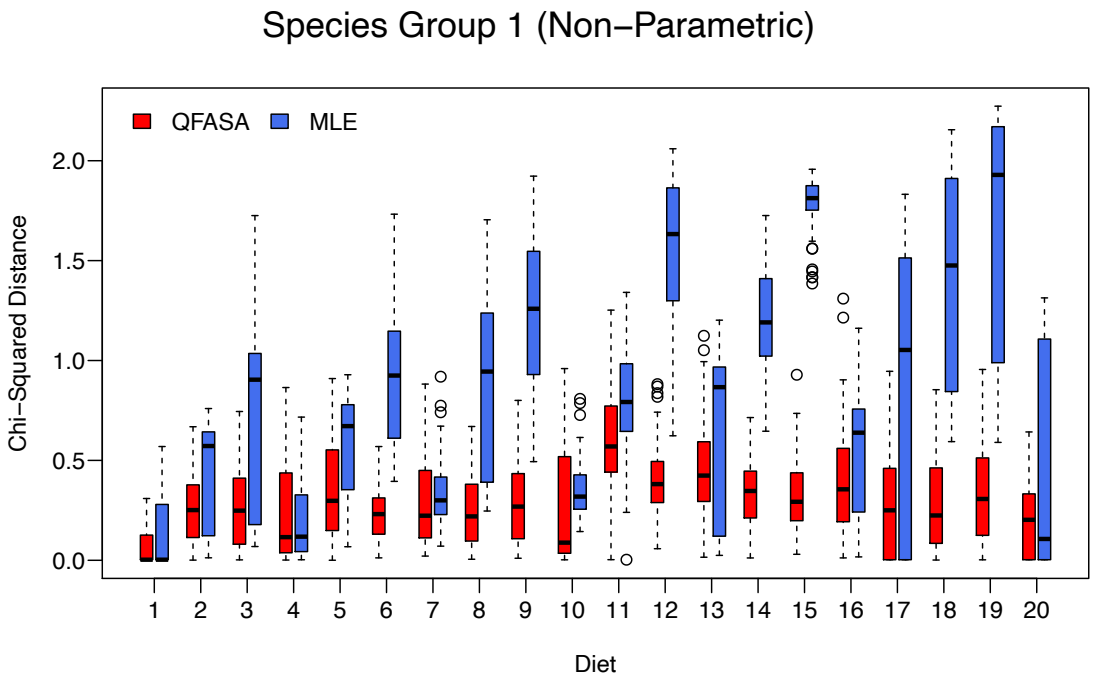

Figure 4.10: Boxplot of Chi-Squared distances for 20 diet simulations with 100 non-parametric pseudo-predators each, on species group 1.

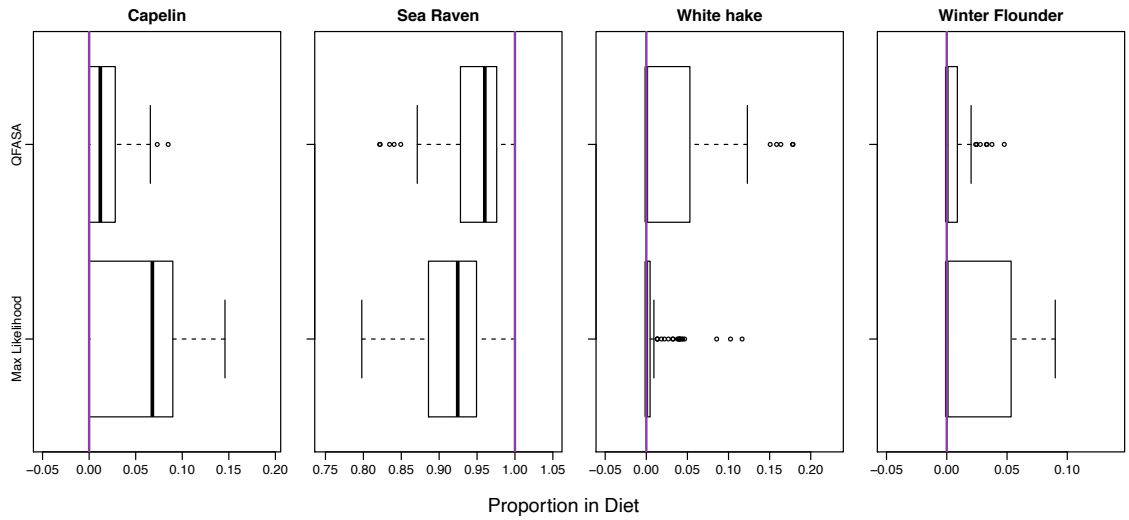

Figure 4.11: Boxplots of estimated diet proportions of 100 non-parametric pseudo-predators using species group 1 and diet 11 for both QFASA and MLE methods, where true diet is shown in purple.

between 0 and 15%. For sea raven, QFASA is underestimating between 0 and 15% and MLE is underestimating between 0 and 20%. For the white hake and winter flounder, both methods have accurate median proportions, but QFASA has larger variability for white hake, and MLE has larger variability for winter flounder. When comparing the FA contributions towards Aitchison's distance to the parametric simulations, these contributions are more equally distributed. The highest contributors (20:4n-6, 16:4n-1, 22:4n-3, and 18:4n-1) account for less than 25%. Therefore, we would expect to see similar results as with our MLE method, as the ilr transformation does not over-weight smaller proportions as Aitchison's distance often does. Despite this being one of the worst estimated diets among the non-parametric simulations, our method is performing very similar to QFASA , and in the case of white hake, even better than QFASA.

One diet that appears to have ML estimates much further from the true diet than QFASA from Figure 4.10 is Diet 19. To see what is happening up close, the boxplot of the diet estimates in this case is displayed in Figure 4.12. In this plot, we can

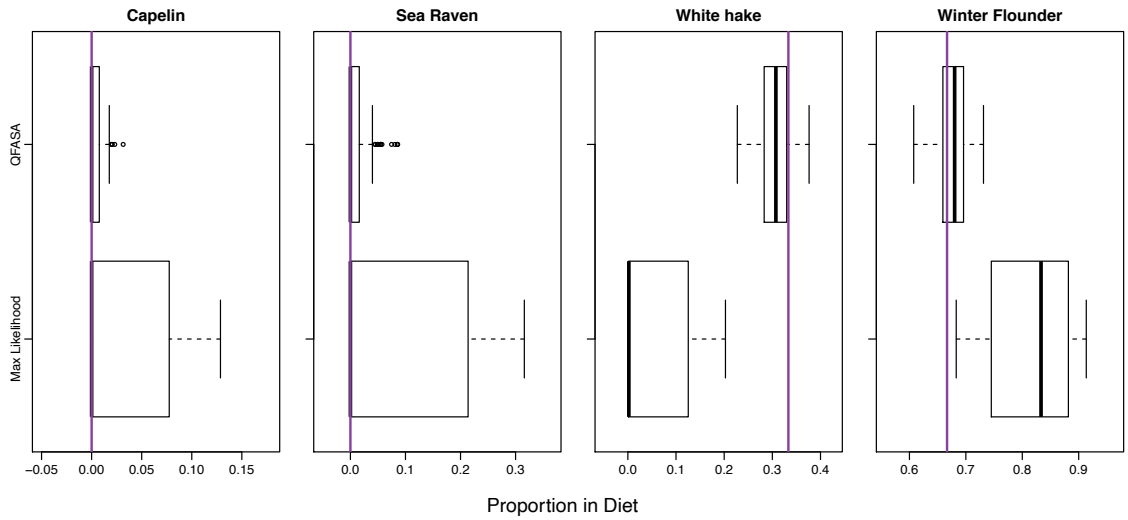

Figure 4.12: Boxplots of estimated diet proportions of 100 non-parametric pseudo-predators using species group 1 and diet 19 for both QFASA and MLE methods, where true diet is shown in purple.

see that for all species, QFASA yields a median estimate nearly exactly on the true proportions, with very little variability. MLE seems to struggle in this case, as the estimates for white hake are between 10-33% underestimated, while those for winter flounder are between 2-25% over estimated. For capelin and sea raven, the median estimate is very close to the true diet, however there is a much larger variability to the estimates than with QFASA. While the pseudo-predators are meant to simulate real life, we can't say which type will be more accurately representative of the FA signature of a predator. For that reason, the most important comparison of this model will be using real life data in Section 6.

In Figure 4.13, we see similar errors as with diet 11, however with much less variability. ML estimates are slightly overestimating capelin by approximately 5%, underestimating sea raven by less than 5%, and has estimated white hake and winter flounder to within 1%. QFASA is performing similarly, with slightly more accurate estimates for capelin and sea raven. Once again, the 4 largest contributing FAs only account for 23%, and thus the weights of the FAs are much more equally distributed than with

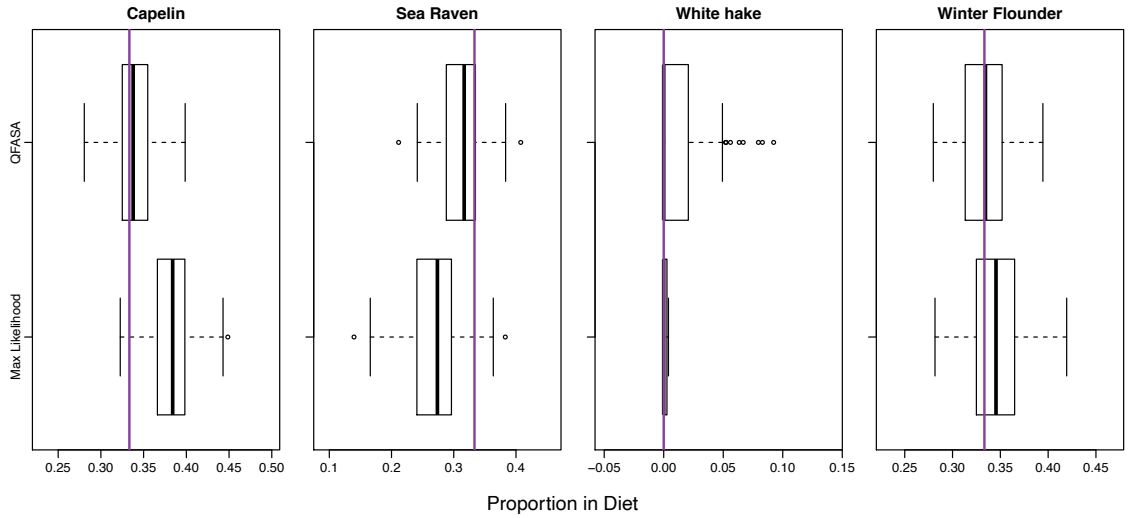

Figure 4.13: Boxplots of estimated diet proportions of 100 non-parametric pseudo-predators using species group 1 and diet 7 for both QFASA and MLE methods, where true diet is shown in purple.

the parametric simulations. QFASA seems to give better estimates when it is not only several FAs are contributing large percentages to Aitchison's distance. In these instances, QFASA and MLE yield similar estimates. Thus, it appears that our MLE is performing well, even when the assumptions about normality are not taken into account.

4.3 Bootstrap Intervals

4.3.1 Bootstrap settings

One of the main goals with this method was to be able to make inference on the true diets. While TMB has the capability to return standard errors of the estimates, there is a gap in the literature about how this is done. As well, there is some debate among fellow statisticians about the accuracy of these standard errors, as well as the optimization of the likelihood when requesting these standard errors. Therefore, until some of these issues are resolved, we will perform parametric bootstraps in order to

obtain marginal confidence bounds around the diet estimates.

To apply this to the simulations, for each diet of species group 1 mentioned in Section 4.2, $n = 10$ pseudo-predators were generated parametrically, and the diet estimated using the same technique used in previous simulations described in Section 3.2. Then, these estimates became the "true" diet, and $n = 10$ pseudo-predators were generated parametrically, and diets were estimated. This last step was repeated 100 times, so that we have 100 estimates of pseudo-predator 1, 100 estimates of pseudo-predator 2, and so on. From this, for all 10 of the pseudo-predators, we obtained 95% marginal confidence bounds, using the 2.5% and 97.5% percentiles.

In general, given initial diet estimates $\hat{\boldsymbol{\alpha}}$ from n predators, you can obtain bootstrap marginal confidence intervals for the diets following these steps:

1. Generate n pseudo-predators using the ML estimates, $\hat{\boldsymbol{\alpha}}$, for the true diet proportions, and the quartered diagonal matrix entries, $\hat{\boldsymbol{\sigma}}$ for the true variance-covariance matrix of the error.

2. Estimate the diet proportions, $\hat{\boldsymbol{\alpha}}^r$, for the n pseudo-predators from step 1.

3. Repeat steps 1 and 2 r times.

4. For each of the n predators, you now have r parametric bootstrap replicates. Obtain the 2.5% and 97.5% quantiles of these replicates for each diet proportion to obtain the marginal CI bounds.

4.3.2 Bootstrap Results

For the simulation results, we have $n = 10$ predators each with $r = 100$ bootstrap replicates. Below, in Table 4.7, the true diet, the ML diet estimate, and the parametric bootstrap confidence intervals are displayed below for pseudo-predator 1 (the first of the $n = 10$; generated parametrically), using the diets displayed in Table 4.4 with species group 1 described in Table 4.3.

It is important to note that all diet proportions are restricted between 0 and 1. Therefore, when the true diet proportion is on these bounds, it is impossible to obtain a confidence bound below 0, or above 1, using the percentile method. To explain this, consider when the true diet proportion is 0.5. We would expect to get some estimates that are less than 0.5, and some estimates that are greater than 0.5, but likely few, if any at all, that are exactly equal to 0.5. So using the percentile method, the lower bound would most likely be less than 0.5, and the upper bound would be greater than 0.5. However, when the true diet is 0, we will have no estimates that are below the true value, and very few, if any, that are exactly equal to the value. Therefore, we will most likely obtain a lower bound for 0 that is greater than 0. Similarly, for 1, we will most likely obtain a upper bound that is less than 1. Rounding the lower and upper bounds can help with this issue, as when rounded to 3 decimal places, most of these confidence bounds on the extremes tend to be exactly the true values.

From Table 4.7, we can see that the ML estimates are fairly accurate and are even exact with a few diets. This aligns with the results seen in Section 4.2. It can be seen that nearly all marginal intervals, rounded to 3 decimals, include the true diet values, with the exceptions being proportions on the boundaries. This is due to the inability to obtain estimates below the lower boundary, or above the upper boundary, thereby restricting the true proportion to be either slightly outside the confidence bounds, or exactly on one of the confidence bounds.

To explore this further, coverage probabilities were obtained using the $n = 10$ pseudo-predators and their replicates. While this is a small sample size, and larger samples could be explored for coverage probability, these were quite lengthy to perform, so this can give us a sense of how these intervals are performing. These coverage probabilities are shown in Table 4.8. Since they are marginal intervals, individual species may have different coverage probabilities. Therefore, we first explored the coverage individually for the species, but also included coverage probability for the entire vector. That is, how many of the $n = 10$ confidence bounds include the entire diet vector.

		Capelin	Sea Raven	White Hake	Winter Flounder
Diet 1	α	1	0	0	0
	$\hat{\alpha}$	1.000	0.000	0.000	0.000
	CI	1.000-1.000	0.000-0.000	0.000-0.000	0.000-0.000
Diet 2	α	0.667	0.333	0	0
	$\hat{\alpha}$	0.661	0.335	0.003	0.000
	CI	0.644-0.688	0.310-0.355	0.000-0.016	0.000-0.000
Diet 3	α	0.667	0	0.333	0
	$\hat{\alpha}$	0.645	0.000	0.355	0.000
	CI	0.617-0.670	0.000-0.041	0.314-0.375	0.000-0.000
Diet 4	α	0.667	0	0	0.333
	$\hat{\alpha}$	0.661	0.000	0.015	0.324
	CI	0.632-0.681	0.000-0.042	0.000-0.048	0.292-0.341
Diet 5	α	0.333	0.667	0	0
	$\hat{\alpha}$	0.325	0.667	0.009	0.000
	CI	0.301-0.349	0.619-0.691	0.000-0.041	0.000-0.000
Diet 6	α	0.333	0.333	0.333	0
	$\hat{\alpha}$	0.325	0.319	0.356	0.000
	CI	0.295-0.351	0.275-0.363	0.324-0.396	0.000-0.000
Diet 7	α	0.333	0.333	0	0.333
	$\hat{\alpha}$	0.332	0.320	0.018	0.330
	CI	0.302-0.360	0.258-0.365	0.001-0.074	0.283-0.354
Diet 8	α	0.333	0	0.667	0
	$\hat{\alpha}$	0.317	0.000	0.683	0.000
	CI	0.297-0.339	0.000-0.035	0.654-0.700	0.000-0.000
Diet 9	α	0.333	0	0.333	0.333
	$\hat{\alpha}$	0.322	0.000	0.354	0.323
	CI	0.290-0.340	0.000-0.054	0.314-0.391	0.293-0.347
Diet 10	α	0.333	0	0	0.667
	$\hat{\alpha}$	0.325	0.000	0.009	0.667
	CI	0.289-0.343	0.000-0.047	0.001-0.049	0.623-0.688
Diet 11	α	0	1	0	0
	$\hat{\alpha}$	0.000	0.989	0.011	0.000
	CI	0.000-0.014	0.956-0.999	0.001-0.040	0.000-0.000
Diet 12	α	0	0.667	0.333	0
	$\hat{\alpha}$	0.000	0.651	0.349	0.000
	CI	0.000-0.000	0.611-0.681	0.319-0.389	0.000-0.000
Diet 13	α	0	0.667	0	0.333
	$\hat{\alpha}$	0.000	0.659	0.011	0.330
	CI	0.000-0.000	0.620-0.685	0.002-0.057	0.297-0.358
Diet 14	α	0	0.333	0.667	0
	$\hat{\alpha}$	0.00	0.321	0.679	0.000
	CI	0.000-0.000	0.287-0.350	0.650-0.713	0.000-0.000
Diet 15	α	0	0.333	0.333	0.333
	$\hat{\alpha}$	0.000	0.320	0.355	0.324
	CI	0.000-0.000	0.282-0.351	0.319-0.403	0.290-0.358
Diet 16	α	0	0.333	0	0.667
	$\hat{\alpha}$	0.000	0.321	0.014	0.665
	CI	0.000-0.000	0.282-0.363	0.000-0.066	0.620-705
Diet 17	α	0	0	1	0
	$\hat{\alpha}$	0.000	0.000	1.000	0.000
	CI	0.000-0.000	0.000-0.000	1.000-1.000	0.000-0.000
Diet 18	α	0	0	0.667	0.333
	$\hat{\alpha}$	0.000	0.000	0.681	0.319
	CI	0.000-0.000	0.000-0.000	0.648-0.710	0.290-0.352
Diet 19	α	0	0	0.333	0.667
	$\hat{\alpha}$	0.000	0.000	0.354	0.646
	CI	0.000-0.000	0.000-0.000	0.317-0.390	0.610-0.683
Diet 20	α	0	0	0	1
	$\hat{\alpha}$	0.000	0.000	0.011	0.989
	CI	0.000-0.000	0.000-0.000	0.001-0.046	0.954-0.999

Table 4.7: True diet, ML estimate, and parametric bootstrap confidence intervals for predator 1, in a sample of $n = 10$ pseudo-predators, for all 20 diets in species group 1.

Diet	Capelin	Sea Raven	White Hake	Winter Flounder	Total
1	**1**	**1**	**1**	**1**	1
2	0.9	0.9	**1**	**1**	0.9
3	0.9	**0.9**	1	**1**	0.8
4	0.8	**1**	**0.5**	0.8	0.4
5	1	1	**0.8**	**1**	0.8
6	1	0.9	0.9	**1**	0.8
7	1	1	**0.3**	0.9	0.3
8	0.9	**0.9**	0.9	**1**	0.8
9	0.9	**0.9**	1	0.8	0.6
10	0.9	**1**	**0.5**	0.8	0.4
11	**1**	0.4	**0.4**	**1**	0.4
12	**1**	1	1	**1**	1
13	**1**	0.9	**0.5**	0.8	0.5
14	**1**	1	1	**1**	1
15	**1**	0.9	0.9	0.8	0.7
16	**1**	0.9	**0.4**	0.8	0.4
17	**1**	**1**	1	**1**	1
18	**1**	**1**	0.8	0.8	0.8
19	**1**	**1**	0.9	0.9	0.9
20	**1**	**1**	**0.3**	**0.3**	0.3
Mean	0.965	0.930	0.755	0.885	0.69

Table 4.8: Coverage probabilities based on $n = 10$ pseudo-predators and marginal 95% percentile bounds based on 100 bootstrap replicates, first for individual prey species, then for the total diet composition. Bold values indicate where true proportions were on the edge of the simplex (0 or 1).

As these are marginal confidence bounds, we could obtain differing coverage probabilities for different species. In particular, we expect to see lower coverage probabilities when the true diet is on, or close to the boundaries. The coverage probability for capelin and sea raven seem to be quite high, averaging 0.965 and 0.930 respectively, close to our confidence level of 95%. When comparing the mean absolute difference in mean ilr transformed FA signatures between the species, Capelin is the most different, making this species more easily distinguishable than the rest, and thus yielded better results. Capelin also had the largest sample size by almost double. White hake and winter flounder seem to exclude the true diet more often, with coverage probabilities averaging 0.755 and 0.855 respectively. While these are 20 and 10% lower than our confidence level, the sample sizes were quite small, and we expect to see larger coverage with larger samples. The average total coverage probability was 0.69. Note that this value requires the bounds for every prey species to capture the true diet proportions simultaneously. This will be lower than individual values as it is a stricter requirement and all of our diets included at least one boundary value (specifically 0). If we had diets completely inside the bounds of the simplex, we would expect these to be higher, and closer to the confidence level.

Taking a closer look at two diets, specifically one with low coverage probability (diet 20) and one with high coverage probability (diet 12), we obtain Figures 4.14 and 4.15 respectively. Once again, we have rounded the values to 3 decimal places. We can see with diet 20, for capelin and sea raven, the true proportions are 0 which are boundary values, and once rounded, they have MLEs right on 0, and no width to the intervals. White hake also has a true proportion of 0, however, the MLEs in this case are slightly positive. Therefore, we get interval values that are always non negative, and are more difficult to capture the true proportion 0. Similarly, with winter flounder, we have another true proportion on the boundary, as 1. As discussed before, our MLEs are restricted to be less than or equal to 1, so it is difficult to capture the true proportion in the interval. As expected, the worst coverage probability in our simulation was for a case with all boundary values in the true diet.

For diet 12, there are two true proportions of 0 for capelin and winter flounder. Both

Computer/Thesis/ErrorSimulations/Bootstraps/D20-bootplotrounded.pdf

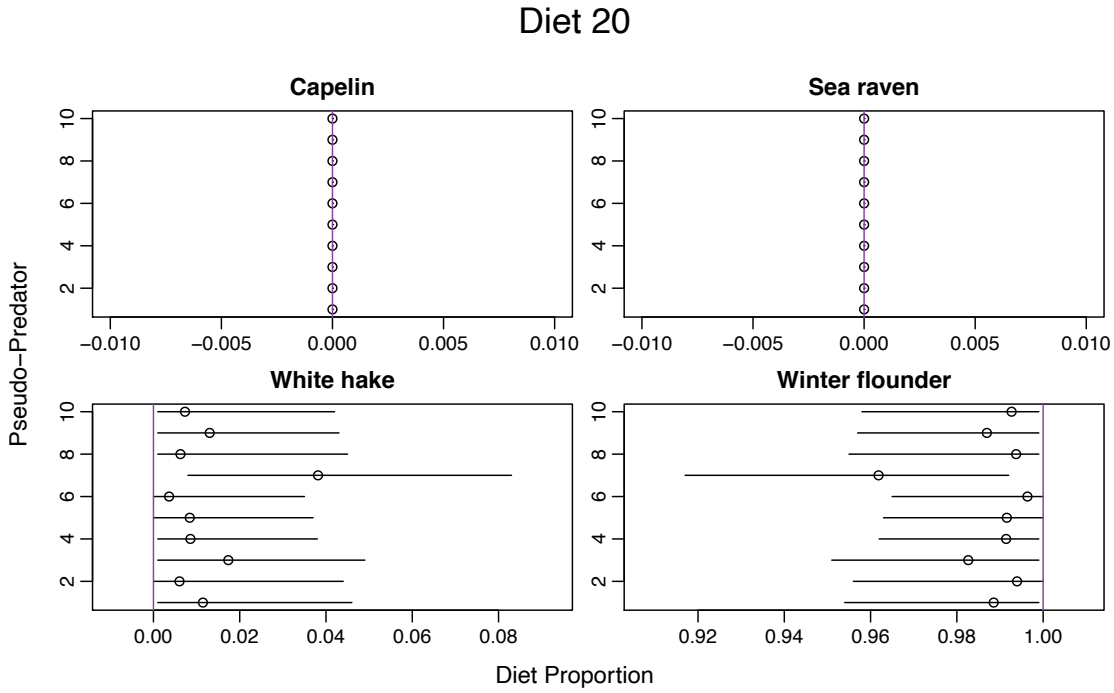

Figure 4.14: Marginal confidence bounds for 10 pseudo-predators with true diet (diet 20) shown in purple, rounded to 3 decimal places. The circle represents the MLE of the diet proportion.

have MLEs which when rounded, are right on 0 with no variability. For sea raven and white hake, which have true proportions of 0.33 and 0.67 respectively, MLE's are no more than 5% away from the true values, with relatively narrow intervals. Despite the narrow intervals, all 10 of the marginal CIs for both white hake and sea raven capture the true proportions. This shows that while there is difficulty on boundary values, the marginal CIs appear to perform well when we have proportions not on the boundaries.

Computer/Thesis/ErrorSimulations/Bootstraps/D12-bootplotrounded.pdf

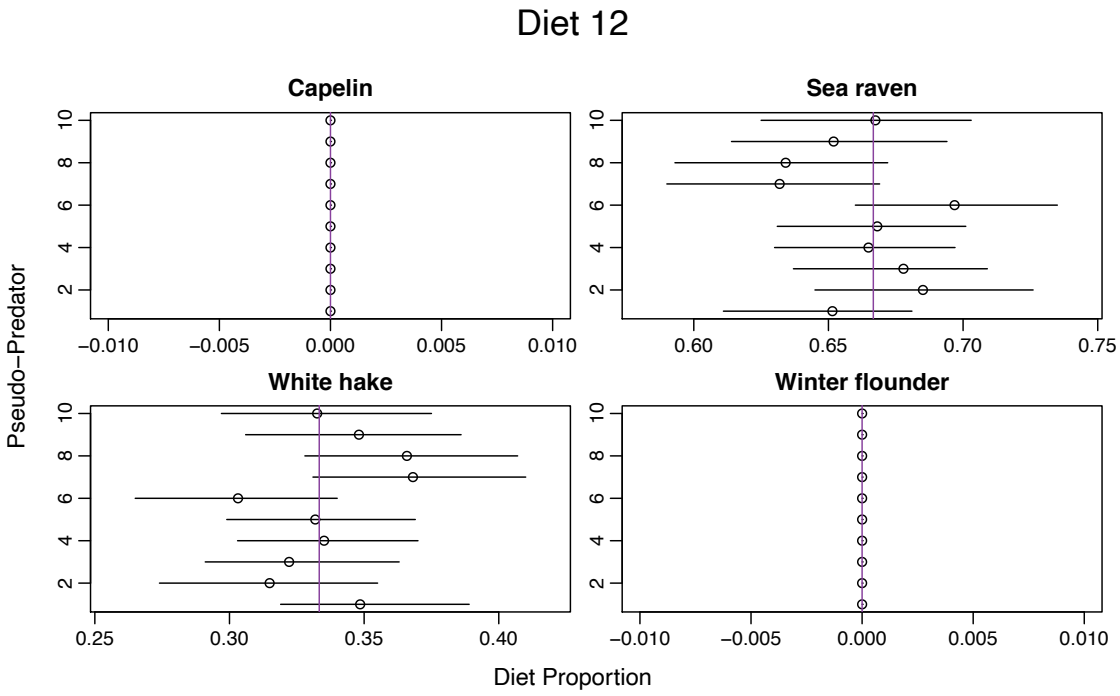

Figure 4.15: Marginal confidence bounds for 10 pseudo-predators with true diet (diet 12) shown in purple, rounded to 3 decimal places. The circle represents the MLE of the diet proportion.

Chapter 5

Covariates

One of the goals of a maximum likelihood approach to QFASA is to include covariates in the estimation of diet. Previously, predator diets were estimated using QFASA, and then modelled using Dirichlet regression or a similar approach (see Section 2.7.4). For our approach, these two steps are performed simultaneously, allowing the uncertainty in diet estimates to be included in the inference on the covariates. Therefore, information is not being lost from estimation to inference. This is a valuable tool for ecologists, as we can estimate diets across different groups, such as those created by sex (male/female), age (pup/adolescent/adult, or a continuous variable), location (specific areas, or a continuous variable such as longitude and latitude), and in the same step, test for a difference among the groups. How this can be performed is described in the next section.

5.1 Methodology

Using a link function, we can estimate the diet composition while including coefficients for any covariates we may be interested in. This link function is the same as that used in Dirichlet regression (Equation 2.17), which is:

$$\alpha_{j1} = \frac{1}{1 + \sum_{s=2}^{I} e^{\mathbf{W}_j \boldsymbol{\beta}_s}}, \qquad \alpha_{ji} = \frac{e^{\mathbf{W}_j \boldsymbol{\beta}_i}}{1 + \sum_{s=2}^{I} e^{\mathbf{W}_j \boldsymbol{\beta}_s}} \qquad i = 2, ..., I, j = 1..., n \tag{5.1}$$

where α_{ji} represents the proportion of prey species i in the j^{th} predator's diet, and $\mathbf{W}$ represents the $n \times (p+1)$ covariate matrix in which the first column is all 1s for the intercepts, and the rest of the matrix is filled with the covariate values, as follows:

$$\mathbf{W} = \begin{bmatrix} 1 & W_{1,2} & W_{1,3} & \cdots & W_{1,p} \\ 1 & W_{2,2} & W_{2,3} & \cdots & W_{2,p} \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ 1 & W_{n,2} & W_{n,3} & \cdots & W_{n,p} \end{bmatrix}$$

where p is the number of covariates and n is the number of predators. The $\boldsymbol{\beta}_i$ in Equation 5.1 represents the vector of coefficients to be optimized over, and i indicates the prey species for which these coefficients represent, and $\mathbf{W}_j$ represents the j^{th} row vector from the matrix above. Therefore, $\boldsymbol{\beta}_i = (\beta_{i,0}, \beta_{i,1}, ..., \beta_{i,p})^T$, where $\beta_{i,0}$ is the intercept coefficient for prey species i, and $\beta_{i,p}$ is the coefficient for covariate p and prey species i.

Using this link function, we have a direct relationship from $\boldsymbol{\beta}$ to $\boldsymbol{\alpha}$. Therefore, we keep the likelihood function from Equation 3.15 in terms of $\boldsymbol{\alpha}$, but the unknown parameters are now $\boldsymbol{\beta}$, so a function is added at the beginning to obtain $\boldsymbol{\alpha}$ from $\boldsymbol{\beta}$. The likelihood is then optimized as before, but now over $\boldsymbol{\beta}$. That is:

$$\begin{aligned} \mathcal{L} &= \int \cdots \int \prod_{j=1}^{n.pred} f(\mathbf{Y}_j | \mathbf{Z}_j, \boldsymbol{\beta}, \mathbf{T}, \mathbf{T}_\epsilon, \mathbf{X}, \mathbf{W}) f(\mathbf{Z}_j) d\mathbf{Z}_{j1} \cdots d\mathbf{Z}_{jI} \\ &= \int \cdots \int \prod_{j=1}^{n.pred} \left(\frac{1}{(2\pi)^{\frac{D-1}{2}} |\boldsymbol{\Sigma}_\epsilon|^{1/2}} \exp\left\{ -\frac{1}{2} (\mathbf{Y}_j - \boldsymbol{\eta}_j^*)^T \boldsymbol{\Sigma}_\epsilon^{-1} (\mathbf{Y}_j - \boldsymbol{\eta}_j^*) \right\} \times \right. \\ &\left. \prod_{i=1}^{I} \frac{1}{(2\pi)^{\frac{D-1}{2}} |\hat{\boldsymbol{\Sigma}}|^{1/2}} \exp\left\{ -\frac{1}{2} (\mathbf{Z}_{ji} - \bar{\mathbf{X}}_i)^T \hat{\boldsymbol{\Sigma}}^{-1} (\mathbf{Z}_{ji} - \bar{\mathbf{X}}_i) \right\} \right) d\mathbf{Z}_{j1} \cdots d\mathbf{Z}_{jI} \end{aligned} \tag{5.2}$$

where $\boldsymbol{\eta}_{oj}^* = \sum_{i=1}^{I} \alpha_{ji} z_{ji}$ as before, $\boldsymbol{\eta}_j^* = \text{ilr}(\boldsymbol{\eta}_{oj}^*)$, and $\boldsymbol{\alpha}_j$ is described by Equation 5.1. Note, on the surface this appears to be the same likelihood as in Equation 3.15, however here, we are optimizing over $\boldsymbol{\beta}$ and we are conditioning on the covariate matrix $\mathbf{W}$. Also note, this assumes that the ilr transformation of the observed ($\mathbf{X}$) and unobserved ($\mathbf{Z}$) prey FA signatures are normally distributed as before. There are several other alternative distributions that could be used instead (see Section 3.2).

This method will yield a summary diet estimate for each unique set of covariates. Therefore, if we consider the simple case with just one covariate (represented by an indicator variable), sex, one diet will be estimated for all the females, and one diet will be estimated for all the males. If we have a continuous covariate, such as sea-surface temperature, the method will yield a diet for each unique sea-surface temperature.

5.2 Simulations

Our goal with the simulations was to determine how well the method performed at estimating a summary diet from two different groups (ie: male/female) in a variety of situations, such as when the diet proportions are similar for all prey species, different, or somewhere in between. Since the simulations from the first method showed that the MLE method is performing well for diets throughout the simplex, the number of diets to choose from was reduced by selecting an increment of $\frac{1}{2}$ in "*make_diet_grid*", which gives 10 different diets, shown in Table 5.2. All unique combinations of these two diets are then used as the "true diets" for the male and female groups. Note that it does not matter which group we label male or female for this simulation, therefore, we do not need to run, for example, diet 1 for male, diet 2 for female, and diet 2 for male and diet 1 for female, as it will yield the same results. This yields 45 different combinations of diets.

In order to obtain $\boldsymbol{\beta}$ from $\boldsymbol{\alpha}$, the inverse link function is required. While this may not yield a unique $\boldsymbol{\beta}$ from $\boldsymbol{\alpha}$ (this function is not one-to-one if there are continuous covariates), it can be seen that zero proportions in $\boldsymbol{\alpha}$ are impossible regardless due to the log ratio in the inverse link function. Therefore, we first use the multiplicative replacement method using $\lambda = 0.00005$, which will set all 0 proportions to 0.00005, and adjust the remaining proportions so that the elements sum to 1. Once the zeros are replaced, we can use the inverse of the link function, shown in Equation 5.3, to find the "true" $\boldsymbol{\beta}$ values.

Diet	Capelin	Sea Raven	White Hake	Winter Flounder
1	1	0	0	0
2	0.5	0.5	0	0
3	0.5	0	0.5	0
4	0.5	0	0	0.5
5	0	1	0	0
6	0	0.5	0.5	0
7	0	0.5	0	0.5
8	0	0	1	0
9	0	0	0.5	0.5
10	0	0	0	1

Table 5.1: Equally spaced diets over the simplex included in the covariate simulations.

$$\boldsymbol{\beta}_i = \mathbf{W}^{-1} \log\left(\frac{\boldsymbol{\alpha}_i}{\boldsymbol{\alpha}_1}\right), i = 2, ..., I \tag{5.3}$$

$\boldsymbol{\alpha}_i$ represents the $n \times 1$ vector of diet proportions in all n predators of the i^{th} prey species, $\mathbf{W}^{-1}$ represents the inverse of the covariate matrix $\mathbf{W}$. Since $\mathbf{W}$ is not necessarily a square matrix, the Moore-Penrose generalized inverse, found using "*ginv*" in R, is used. If all covariates are categorical in nature, and are represented by indicator variables, there is a one-to-one function between $\boldsymbol{\alpha}$ and $\boldsymbol{\beta}$ and we could solve for these equations by hand. We have done just that for the one indicator variable case in Section 5.4. However, for more complex situations with more prey species and more covariates, there is not a one-to-one function between $\boldsymbol{\alpha}$ and $\boldsymbol{\beta}$ as there will be many more diet proportions than coefficients. Since the use of the generalized inverse has yielded exactly the same results in all attempted examples when using one covariate, we are justified in assuming it will yield valid results for cases with more covariates, and will use this function to obtain starting coefficient values.

To obtain the "true" diets, $\boldsymbol{\alpha}$, for $n = 10$ males and $n = 10$ females, we first create a covariate matrix $\mathbf{W}$ of dimension 20×2 with first column $\mathbf{1}$ and second column values equal to 1 for the first 10 rows, representing the males, and values of 0 for the last 10 rows, representing the females. This covariate matrix is then substituted into the link functions in Equation 5.1, then multiplied by the "true" $\boldsymbol{\beta}$ values to obtain the true diet for the male group, and the true diet for the female group. 10 male

pseudo-predators are then generated using the parametric technique described in 4.2 and the first "true" diet, and 10 female pseudo-predators are generated using the same technique and the second "true" diet. Then, using the procedure outlined in Section 5.1, the MLE method with covariates is run to obtain estimates of $\boldsymbol{\beta}$, which in turn give the diet estimates, $\hat{\boldsymbol{\alpha}}$. This procedure is then repeated 50 times, so that we have 50 summary diets for all females and all males, for sample sizes of $n = 10$ in each group.

5.3 Results

Similar to the analysis of simulation results in Section 4.2, chi-squared distances (Definition 2.19) were calculated between the ML diet estimates and the true diet, and similarly for QFASA estimates. For all 45 combinations of unique diets simulated, 50 estimates were obtained from 10 male and 10 female pseudo-predators. The mean chi-squared distance of all 50 estimates for each combination is recorded in Table 5.2. In most cases, the mean (Euclidean mean over each prey species) QFASA estimates are further from the true diets than the MLEs, however, in some cases, they are comparable (diet 1 estimates). It can be seen that no matter what the diet is for the other group, the estimates appear to be similar distances away from the true diet when the same diet is used. For example, for all groups that have diet 1, the ML estimates are between 0.04 and 0.05 away from true values, on average, for all diets in the other group. Thus, the accuracy of group 1's estimation does not depend on group 2's diet.

In order to visualize this, these mean values are plotted in Figure 5.1. From this plot, we can clearly see that QFASA estimates tend to be further away from the true diet for both groups. For the few instances where this isn't true, the ML estimates are only slightly higher, by only 0.01 or 0.02. This only occurs when estimating diet 1 (1, 0, 0, 0) and diet 8 (0, 0, 1, 0). Because these diet estimates are on the boundaries of the simplex and estimates are restricted by these bounds, we will only ever get estimates on the inside of the bounds, making it less likely that the estimated mean diet proportions are close to the truth. The MLE diet estimates for both groups are

	Males			Females		
Combo	Diet	MLE	QFASA	Diet	MLE	QFASA
1	**1**	**0.05**	**0.03**	**2**	**0.10**	**0.24**
2	1	0.04	0.03	3	0.12	0.36
3	1	0.05	0.03	4	0.15	1.31
4	1	0.05	0.03	5	0.17	0.85
5	1	0.05	0.03	6	0.11	0.45
6	1	0.05	0.03	7	0.15	1.60
7	1	0.04	0.03	8	0.04	0.03
8	1	0.04	0.03	9	0.07	0.62
9	1	0.04	0.03	10	0.10	1.09
10	2	0.13	0.25	3	0.13	0.37
11	2	0.08	0.24	4	0.15	1.31
12	2	0.09	0.25	5	0.18	0.86
13	2	0.09	0.23	6	0.11	0.46
14	2	0.09	0.24	7	0.17	1.60
15	2	0.11	0.24	8	0.10	0.03
16	2	0.09	0.23	9	0.09	0.61
17	2	0.09	0.25	10	0.12	1.09
18	3	0.14	0.38	4	0.18	1.33
19	3	0.17	0.39	5	0.16	0.84
20	3	0.13	0.38	6	0.13	0.46
21	3	0.15	0.37	7	0.17	1.60
22	3	0.12	0.37	8	0.09	0.03
23	3	0.12	0.37	9	0.09	0.61
24	3	0.12	0.38	10	0.11	1.09
25	4	0.13	1.30	5	0.15	0.85
26	4	0.16	1.31	6	0.11	0.47
27	**4**	**0.18**	**1.31**	**7**	**0.18**	**1.60**
28	4	0.15	1.32	8	0.07	0.03
29	4	0.14	1.31	9	0.07	0.61
30	4	0.14	1.31	10	0.10	1.09
31	5	0.15	0.85	6	0.09	0.46
32	5	0.16	0.85	7	0.16	1.60
33	5	0.17	0.86	8	0.09	0.03
34	5	0.14	0.85	9	0.08	0.62
35	5	0.16	0.86	10	0.13	1.09
36	6	0.12	0.46	7	0.16	1.60
37	6	0.11	0.47	8	0.09	0.03
38	6	0.12	0.46	9	0.10	0.62
39	6	0.11	0.46	10	0.11	1.09
40	7	0.14	1.60	8	0.10	0.03
41	7	0.14	1.60	9	0.10	0.62
42	7	0.14	1.60	10	0.12	1.09
43	8	0.09	0.03	9	0.07	0.62
44	8	0.09	0.03	10	0.14	1.09
45	9	0.04	0.62	10	0.11	1.09

Table 5.2: Mean chi-squared distances between true diet and summary diet estimates for two groups (males and females) with two unique diets, for sample size $n = 10$ and replicates $r = 50$.

Computer/Thesis/ErrorSimulations/Covariates/covchisquareddist.pdf

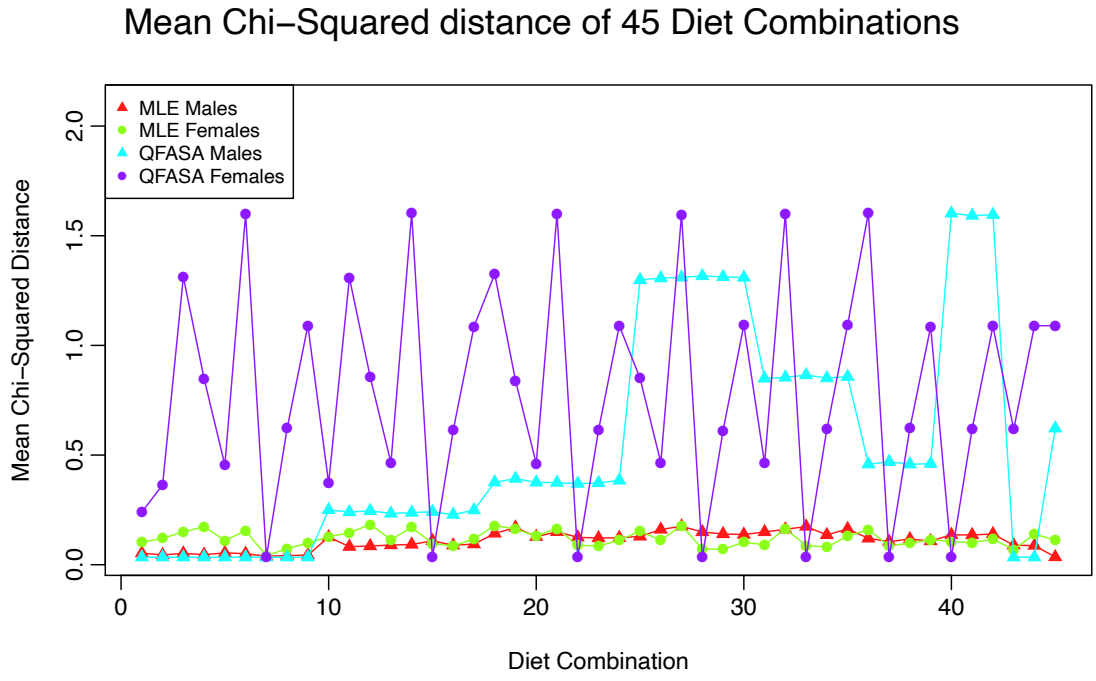

Figure 5.1: Mean chi-squared distances between the estimated and true diets of males and females, using QFASA and MLE, for the diets displayed in Table 5.2.

less than a chi-squared distance of 0.2 away from the true diet on average, whereas QFASA estimates reach above 1.5. As mentioned earlier, the estimates for one group are not significantly affected by the other group, if at all, and thus we see patterns in the plot corresponding to the pattern of true diets. For example, all female estimates from QFASA with true diet 7 are an average of approximately 1.60 away from the true diet. These are all the highest peaks for QFASA female estimates on the plot, and due to the order of the diets, there is a cyclic pattern. Similarly, for males, the order has all diet 1s estimated together, then all diet 2s, and so on, so it appears there is a stepping pattern. All in all, there seems to be a significant improvement when estimating a summary diet for groups when using the MLE method with covariates, compared to individually estimating diets using QFASA, and averaging.

Selecting a good case (combination 1; Figure 5.2) where the distances for males and females using QFASA and ML method are low, and a less accurate case (combination

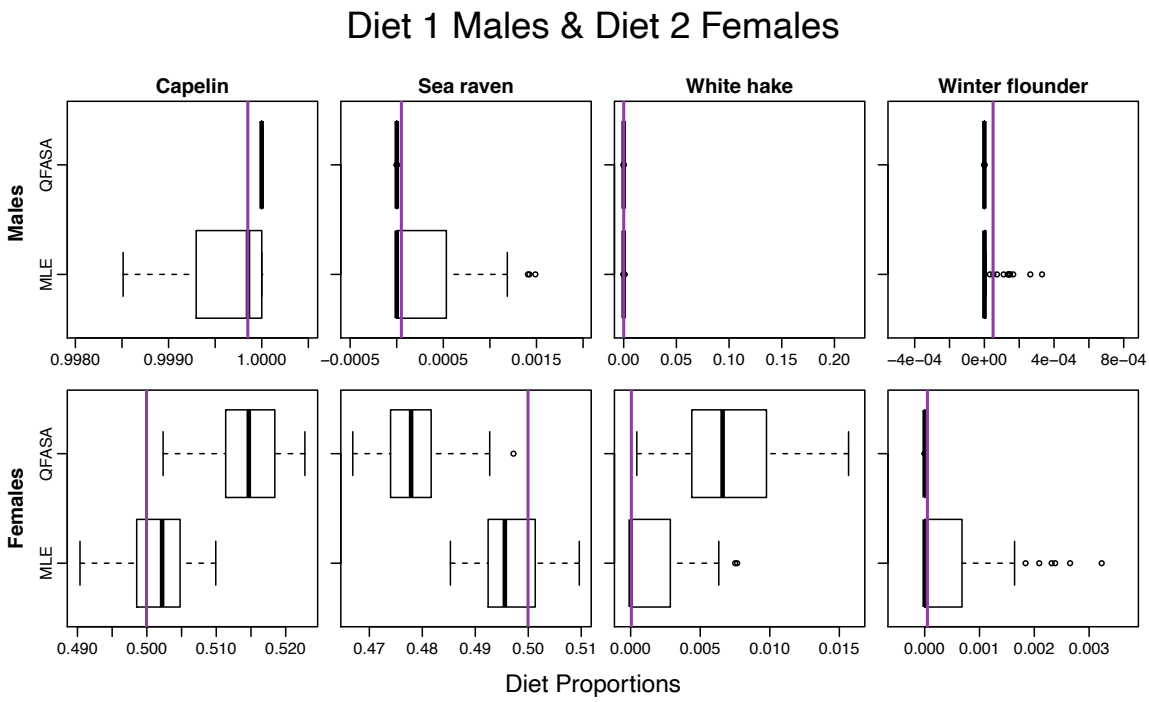

Figure 5.2: Boxplot of summary diets for males (diet 1) and females (diet 2) using $n = 10$ males and $n = 10$ females, and $r = 50$ replicates.

27; Figure 5.3) where the distances are relatively large, we can take a closer look at what is happening with the estimates. Looking at Figure 5.2, we can see that for both males and females, for all 4 species, the median summary estimates using ML method are much closer to the true values than the median QFASA estimate. For males and females and all species, the worst summary estimate that the ML method provides is only 0.015 (1.5%) away from the true proportion. The median estimate for QFASA is often further away from the true proportion than the worst ML estimate (capelin - females, sea raven - females). Even as one of the better estimates for both techniques, it is still clear that the ML covariate method is performing much better than QFASA.

Looking at Figure 5.3, for both sexes and nearly every prey species (with the exception of capelin for females) the ML covariate method is significantly closer to the true proportion than QFASA. While this is one of the worst estimates for both techniques, the MLEs are still nearly exactly on the true proportions, with very little variability. In comparison, QFASA is much further away, and with generally higher variability,

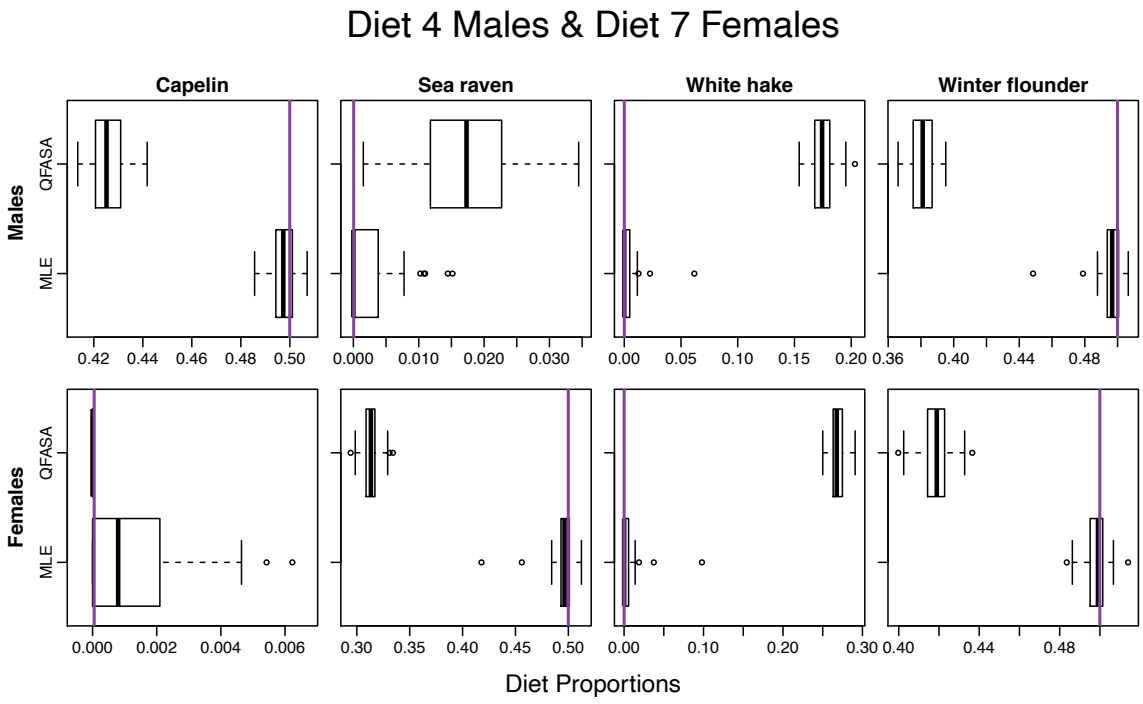

Figure 5.3: Boxplot of summary diets for males (diet 4) and females (diet 7) using $n = 10$ males and $n = 10$ females, and $r = 50$ replicates.

being as far as 0.3 (30%) away from the true proportion (white hake, females). The addition of covariates into the model seems to greatly improve the ability to estimate a summary diet for a group of predators, and also allow us to test for significance, which we will see in the next section.

5.4 Inference

We have now successfully added covariates into our maximum likelihood diet model. An added purpose of this model is to test for differences in diets among different groups, for example, males and females. Using the technique used in Section 5.1, a summary diet can be estimated for the groups. Now, we will discuss how to determine if these summary diets are significantly different from each other.

Let's first consider the simple case of two groups, for example, males and females, and 4 prey species. We would only have one covariate W which takes on 0 if the predator is male, 1 if the predator is female. Thus, the covariate vector corresponding to each male would be $\mathbf{W}_m = (1, 0)$, and the covariate vector corresponding to each female would be $\mathbf{W}_f = (1, 1)$, where the first entry will always be 1 to correspond with the intercept. $\boldsymbol{\beta}$ would then be a 2×3 matrix as follows:

$$\boldsymbol{\beta} = \begin{bmatrix} \boldsymbol{\beta}_2 & \boldsymbol{\beta}_3 & \boldsymbol{\beta}_4 \end{bmatrix} = \begin{bmatrix} \beta_{2,0} & \beta_{3,0} & \beta_{4,0} \\ \beta_{2,1} & \beta_{3,1} & \beta_{4,1} \end{bmatrix}$$

The link functions would simplify to:
Male:

$$\begin{aligned} \alpha_{m1} &= \frac{1}{1 + \sum_{s=2}^{I} e^{\mathbf{W}_m \boldsymbol{\beta}_s}} \\ &= \frac{1}{1 + \sum_{s=2}^{I} e^{\beta_{s,0}}} \end{aligned}$$

$$\begin{aligned}\alpha_{mi} &= \frac{e^{\mathbf{W}_m\boldsymbol{\beta}_i}}{1+\sum_{s=2}^{I} e^{\mathbf{W}\boldsymbol{\beta}_s}} \\ &= \frac{e^{\beta_{i,0}}}{1+\sum_{s=2}^{I} e^{\beta_{s,0}}}\end{aligned}$$

Female:

$$\begin{aligned}\alpha_{f1} &= \frac{1}{1+\sum_{s=2}^{I} e^{\mathbf{W}_f\boldsymbol{\beta}_s}} \\ &= \frac{1}{1+\sum_{s=2}^{I} e^{\beta_{s,0}+\beta_{s,1}}}\end{aligned}$$

$$\begin{aligned}\alpha_{fi} &= \frac{e^{\mathbf{W}_f\boldsymbol{\beta}_i}}{1+\sum_{s=2}^{I} e^{\mathbf{W}\boldsymbol{\beta}_s}} \\ &= \frac{e^{\beta_{i,0}+\beta_{i,1}}}{1+\sum_{s=2}^{I} e^{\beta_{s,0}+\beta_{s,1}}}\end{aligned}$$

where $i = 2, ..., I$, and $I = 4$ is the number of prey species.

We can then solve this system of equations to solve for $\boldsymbol{\beta}$ in terms of $\boldsymbol{\alpha}$. Since our one covariate is an indicator variable, there are the same number of unknowns in both parameters which directly and uniquely determine each other, thus there is a one-to-one function between $\boldsymbol{\beta}$ and $\boldsymbol{\alpha}$. The equations for $\boldsymbol{\beta}$ are shown below:

$$\begin{aligned}\beta_{i,0} &= \log\left(\frac{\alpha_{mi}}{\alpha_{m1}}\right) \\ &= \log\left(\frac{\alpha_{mi}}{1-\alpha_{m2}-\alpha_{m3}-\alpha_{m4}}\right)\end{aligned}$$

$$\begin{aligned}\beta_{i,1} &= \log\left(\frac{\alpha_{fi}}{\alpha_{f1}}\right) - \log\left(\frac{\alpha_{mi}}{\alpha_{m1}}\right) \\ &= \log\left(\frac{\alpha_{fi}}{1-\alpha_{f2}-\alpha_{f3}-\alpha_{f4}}\right) - \log\left(\frac{\alpha_{mi}}{1-\alpha_{m2}-\alpha_{m3}-\alpha_{m4}}\right)\end{aligned}$$

So, if we wish to test for a difference between the diets of males and females, we want to test:

$$\begin{aligned}H_o &: \boldsymbol{\alpha}_m = \boldsymbol{\alpha}_f \\ H_a &: \boldsymbol{\alpha}_m \neq \boldsymbol{\alpha}_f\end{aligned}$$

or

$$\begin{aligned}H_o &: \alpha_{m2} = \alpha_{f2} \text{ and } \alpha_{m3} = \alpha_{f3} \text{ and } \alpha_{m4} = \alpha_{f4} \\ H_o &: \text{At least one } \alpha_{mi} \neq \alpha_{fi} \text{ for } i \in \{2,3,4\}\end{aligned}$$

Thus, we need only one α_{mi} to differ from α_{fi} in order for us to reject our null hypothesis. With our inverse link function from $\boldsymbol{\alpha}$ to $\boldsymbol{\beta}$, we could alternatively write these hypothesis in terms of $\boldsymbol{\beta}$. So we need to determine these equivalent hypothesis. We know that $\boldsymbol{\alpha}_m = \boldsymbol{\alpha}_f$ if and only if $\beta_{2,i} = \beta_{3,i} = \beta_{4,i} = 0$. So, our null and alternate hypotheses become:

$$H_0 : \beta_{2,1} = \beta_{3,1} = \beta_{4,1} = 0, \qquad H_a : \text{ At least one } \beta_{i,1} \neq 0, i \in \{2,3,4\} \quad (5.4)$$

or equivalently:

$$H_0 : [\beta_{2,1}, \beta_{3,1}, \beta_{4,1}] = [0,0,0], \qquad H_a : [\beta_{2,1}, \beta_{3,1}, \beta_{4,1}] \neq [0,0,0] \quad (5.5)$$

Olive (2016) proposes a method to test such a hypothesis based on bootstraps and the Mahalanobis distance (Definition 2.17). Their method to test the more general hypotheses $H_o : \boldsymbol{\mu} = \mathbf{c}$ versus $H_a : \boldsymbol{\mu} \neq \mathbf{c}$ uses the test statistic:

$$T = \hat{\boldsymbol{\mu}} - \mathbf{c}$$

Then, r bootstrap replicates of $\hat{\boldsymbol{\mu}}$ are created and this test statistic is calculated for each bootstrap, T_r^*. The squared Mahalanobis distance (the square of Definition 2.17) between the bootstrap test statistic, and the mean test statistic of the replicates ($\bar{T}^*$) is then calculated for all of the bootstrap replicates $\hat{\boldsymbol{\mu}}^r$, using the sample variance-covariance matrix of the bootstrap test statistics. That is:

$$D_r^2 = D_r^2(\bar{T}^*, S_T^*) = (T_r^* - \bar{T}^*)^T [S_T^*]^{-1} (T_r^* - \bar{T}^*)$$

where $\bar{T}^* = \frac{1}{r}\sum_{i=1}^r T_r^*$ and $S_T^* = \frac{1}{r-1}\sum_{i=1}^r (T_r^* - \bar{T}^*)^T (T_r^* - \bar{T}^*)$. Then, using percentiles of the distances, and a significance level α, we get the closed interval $[0, D_{1-\alpha}]$, where $D_{1-\alpha}$ represents the $(1-\alpha)^{th}$ percentile of the squared Mahalanobis distances, D_r^2. Let $D_0^2 = \bar{T}^{*T} [S_T^*]^{-1} \bar{T}^*$. We will reject H_o for H_a if $D_0 > D_{1-\alpha}$, and fail to reject H_o for H_a otherwise.

To apply this to our application, first, we need r bootstrap replicates of $\hat{\boldsymbol{\beta}}$. So, similar to Section 4.3, we obtain the initial estimate of $\boldsymbol{\beta}$, $\hat{\boldsymbol{\beta}}$, using the methodology in Section 5.1, then using the one-to-one link function from $\boldsymbol{\beta}$ to $\boldsymbol{\alpha}$, we obtain diet estimates, $\hat{\boldsymbol{\alpha}}$ from which to parametrically generate r samples of size n. For each sample, the methodology in Section 5.1 is performed to obtain r bootstrap estimates, $\hat{\boldsymbol{\beta}}^r$.

Since we are testing against the zero vector, our test statistic will simply be:

$$T = \hat{\boldsymbol{\beta}}$$

We can then obtain T_r^* from the r bootstrap replicates $\hat{\boldsymbol{\beta}}^r$, and then get the mean $\bar{T}^*$ and the variance-covariance matrix S_T^*. Using these, we can get the squared Mahalanobis distances D_r^2 as well as D_0^2. After choosing our significance level α, we find the $(1-\alpha)^{th}$ percentile of our bootstrap distances, $D_{1-\alpha}$ and reject H_0 for H_a if

	Diet 1	Diet 2	Chi-squared Distance
No Effect	2	2	0
Small Effect	2	5	2.31
Large Effect	2	8	4.00

Table 5.3: Effect size, diet combination, and chi-squared distance between the true diets used to assess the bootstrap inference method.

$D_0 > D_{1-\alpha}$.

Simulation

In order to determine how the percentile method proposed by Olive (2016) was behaving with our ML methodology, a simulation study was designed. First, we wanted to see how this method performs in a simple case with just one indicator covariate, for example sex (male/female). Since there is a relationship between $\boldsymbol{\alpha}$, the diet estimates, and $\boldsymbol{\beta}$, the model coefficients, we based our effect size on the true diets, $\boldsymbol{\alpha}$.

The chi-squared distance (described in Definition 2.19) between the true diet for males and that for females, was used as the effect size. Using the same 10 diets in Table 5.2 with the 0 proportions replaced with the multiplicative replacement method (Definition 2.14) and $\lambda = 0.00005$ (since the covariate method does not allow for 0 diet proportions), we found three diet combinations: one with no effect, one with a small effect, and one with a large effect. These are displayed in Table 5.3.

To ensure more complex scenarios will yield accurate results as well, a simulation was designed with one covariate, 3 levels (for example, pup, adolescent, adult). Similar to the two group case, chi-squared distance was used to determine effect size, only now we average the distance between each pair. Since there are more groups, there are more options for effect sizes. These are described in Table 5.4.

For this model with one covariate, three groups, we will need two indicator variables. The first, w_1 will be 1 if group 2 (adolescent), 0 otherwise, and the second, w_2 will be 1 if group 3 (adolescent), 0 otherwise. Thus, we will have link functions as follows for $\boldsymbol{\alpha}_1$, the summary diet for group 1 (pups), $\boldsymbol{\alpha}_2$, the summary diet for group 2

	Diet 1	Diet 2	Diet 3	Avg Chi-sq
No Effect	2	2	2	0
One Small Effect	2	2	5	1.54
One Large Effect	2	2	8	2.67
All Small Effect	2	6	8	3.05
All Large Effect	2	8	10	4.00

Table 5.4: Effect size, diet combination, and chi-squared distance between true diets used to assess the bootstrap inference method with 3 groups.

(adolescents), and $\boldsymbol{\alpha}_3$, the summary diet for group 3 (adults):

$$\begin{aligned}\alpha_{11} &= \frac{1}{1+\sum_{s=1}^{4} e^{\mathbf{W}_1\boldsymbol{\beta}_s}} \\ &= \frac{1}{1+\sum_{s=1}^{4} e^{\beta_{s,0}}}\end{aligned}$$

$$\begin{aligned}\alpha_{1i} &= \frac{e^{\mathbf{W}_1\boldsymbol{\beta}_i}}{1+\sum_{s=1}^{4} e^{\mathbf{W}_1\boldsymbol{\beta}_s}} \\ &= \frac{e^{\beta_{i,0}}}{1+\sum_{s=1}^{4} e^{\beta_{s,0}}}\end{aligned}$$

$$\begin{aligned}\alpha_{21} &= \frac{1}{1+\sum_{s=1}^{4} e^{\mathbf{W}_2\boldsymbol{\beta}_s}} \\ &= \frac{1}{1+\sum_{s=1}^{4} e^{\beta_{s,0}+\beta_{s,1}}}\end{aligned}$$

$$\begin{aligned}\alpha_{2i} &= \frac{e^{\mathbf{W}_2\boldsymbol{\beta}_i}}{1+\sum_{s=1}^{4} e^{\mathbf{W}_2\boldsymbol{\beta}_s}} \\ &= \frac{e^{\beta_{i,0}+\beta_{i,1}}}{1+\sum_{s=1}^{4} e^{\beta_{s,0}+\beta_{s,1}}}\end{aligned}$$

$$\alpha_{31} = \frac{1}{1 + \sum_{s=1}^{4} e^{\mathbf{W}_3 \boldsymbol{\beta}_s}} = \frac{1}{1 + \sum_{s=1}^{4} e^{\beta_{s,0} + \beta_{s,2}}}$$

$$\alpha_{3i} = \frac{e^{\mathbf{W}_3 \boldsymbol{\beta}_i}}{1 + \sum_{s=1}^{4} e^{\mathbf{W}_3 \boldsymbol{\beta}_s}} = \frac{e^{\beta_{i,0} + \beta_{i,2}}}{1 + \sum_{s=1}^{4} e^{\beta_{s,0} + \beta_{s,2}}}$$

Therefore, to test if all the diets are the same, we want to test that:

$$H_0 : [\beta_{2,1}, \beta_{3,1}, \beta_{4,1}, \beta_{2,2}, \beta_{3,2}, \beta_{4,2}] = [0, 0, 0, 0, 0, 0]$$

$$H_a : [\beta_{2,1}, \beta_{3,1}, \beta_{4,1}, \beta_{2,2}, \beta_{3,2}, \beta_{4,2}] \neq [0, 0, 0, 0, 0, 0]$$

So, we can use the same technique described in Section 5.4, only extend the $\boldsymbol{\beta}$ vector to include the three additional coefficients. Bootstrapping would be performed in the same way, as would the test statistic, and decision rule.

Results

First, we consider the simple case where we have two groups, say males and females, with 4 prey species (we used species group 1 from Table 4.3). Using the cases described in Table 5.3, the observed and critical distances, D_0 and $D_{1-\alpha}$, are displayed.

We can see that for all of the listed significance levels, α, when there is no effect, we fail to reject H_0 for H_a, correctly concluding that there is no significant difference in summary diets for the male and female groups. For all significance levels in the presence of a small or a large effect, we rejected H_0 for H_a, correctly concluding that there is a significant difference in the summary diets for males and females. Ideally, we would like to repeat this to determine the relationship between effect size and power, however due to limited time, this will be a future endeavour.

	D_0	α	$D_{1-\alpha}$	Reject H_0?
No Effect	0.904	0.01	3.410	No
		0.05	2.649	No
		0.10	2.327	No
Small Effect	11.837	0.01	4.696	Yes
		0.05	3.755	Yes
		0.10	2.843	Yes
Large Effect	30.952	0.01	6.032	Yes
		0.05	3.326	Yes
		0.10	2.269	Yes

Table 5.5: Observed Mahalanobis distances under the null hypothesis with varying effect sizes,the percentile distances with varying significance levels, and decisions for the tests.

Similarly, in Table 5.6, we can see the percentile distances and the distance under the null hypothesis for all the cases described in Table 5.4. Once again, this methodology makes the correct decision in all of the cases we explored; failing to reject H_0 when there is no effect, and rejecting H_0 for H_a, when there is. Even when there is only one of the diets that differs from the other two, this methodology picks up on the effect. This simulation depicts the usefulness and accuracy of this method in detecting differences in diets among differing groups.

This chapter showed how covariates could be included into our novel maximum likelihood approach to diet estimation. Through the use of simulations, it was shown that the covariates allowed accurate and precise estimation of summary diets for each unique combination of covariate values. We also described a method for performing inference on the coefficients of these covariates, thus determining if there is a significant difference among diets of several groups. This is a very important improvement upon the original QFASA method and is an excellent base to build upon such techniques for analysis.

	D_0	α	$D_{1-\alpha}$	Reject H_0?
No Effect	0.823	0.01	5.648	No
		0.05	3.357	No
		0.10	2.327	No
One Small Effect	13.807	0.01	6.335	Yes
		0.05	3.716	Yes
		0.10	3.482	Yes
One Large Effect	29.876	0.01	5.942	Yes
		0.05	3.757	Yes
		0.10	3.200	Yes
All Small Effect	32.053	0.01	7.094	Yes
		0.05	4.638	Yes
		0.10	3.318	Yes
All Large Effect	35.425	0.01	7.094	Yes
		0.05	4.638	Yes
		0.10	3.318	Yes

Table 5.6: Observed Mahalanobis distances under the null hypothesis with varying effect sizes,the percentile distances with varying significance levels, and decisions for the tests.

Chapter 6

Real Life Data

6.1 Data Collection

The empirical dataset used is from a captive feeding study conducted at the Vancouver Aquarium by Chad Nordstrom et al. as described in Nordstrom et al. (2008). The study, conducted between August 28 and October 9, 2003, used 21 harbour seals (Phoca vitulina richardsi) that were recovered from the coastline of British Columbia, Canada by the Vancouver Aquarium's Marine Mammal Rescue Centre staff, or were brought to the rescue facility by members of the public. All seals brought to the facility were unweaned, and estimated to be less than 15 days of age; there was no data on their feeding history. Following arrival at the facility, seals were housed in individual tubs and were tube-fed a homogenous mixture of pure salmon oil (commercial blend), ground Pacific herring (*Clupea pallasii*) and water at a ratio of 3:6:8 by weight for 5-21 days. Seals were then transitioned from the homogenate onto whole herring over a period of 5-6 days after which they received solely herring until the start of the experimental period (4-30 days). Seals were transferred to larger shared pools as they increased in size, during which time every effort was made to feed seals individually.

At the start of the experimental period (Day 0), the seals were placed in one of three diet treatments; only Pacific herring for 42 days, only surf smelt (*Hypomesus pretiosus*) for 42 days, or surf smelt for 21 days followed by herring for 21 days. Because this last diet with both smelt and herring being consumed is fairly complicated, we have excluded this set from our study. Daily food intake (to the nearest 0.01 kg) was recorded for all individuals throughout the study. Individual whole prey were randomly subsampled throughout the study period and stored in airtight plastic bags frozen at -20°C until analysis. In addition to the 3 prey items fed to the seals (salmon oil, herring and smelt), whole individuals were obtained for 8 other species of

fish and invertebrates and stored in airtight plastic bags frozen at -20°C until analysis.

The seals were weighted to the nearest 0.1 kg and a full depth blubber biopsy was obtained at Day 0, Day 21 and Day 42 as described in Nordstrom et al. (2008). Not all seals were available for biopsy at Day 42. Four seals in the herring diet group were released without bipsy 34 days after the start of the experimental diets (after completing the rescue program) resulting in a reduced sample size for this group. The samples were then weighed and placed in a chlorogorm solution with 0.01% butylated hydroxytoluene (BHT) and stored at -30°C.

To obtain species-specific calibration coefficients, full depth blubber biopsies were taken as described in Nordstrom et al. (2008) from 4 captive subadult harbour seals (3 males, 1 female) housed at the Vancouver Aquarium. These individuals had exclusively eaten herring from a single lot (the same lot used in the experimental diet treatments above) for >1 year period prior to the biopsies.

The individual fish and invertebrates were thawed, and fork length was measured to the nearest 0.1 cm; body mass was measured to the nearest 0.1 g. Each individual was then homogenized in a food processor. To determine fat content, lipids were quantitatively recovered in duplicate from samples of the homogenised prey using a modified Folche method (Folch et al. (1957), Iverson et al. (2001)). Lipids were extracted from all blubber samples using a modified Folch method (see Budge et al. (2006)). FA methyl esters (FAME) were prepared from the extracted lipids using an acidic catalyst (the Hilditch method, see Budge et al. (2006)). FAME were analysed in duplicate using temperature-programmed gas liquid chromatography according to Iverson et al. (1997). Individual FAs were reported as mass percent of total FAs. FAs are described by the shorthand nomenclature of A:Bn-X where A represents the carbon chain length, B, the number of double bonds, and X, the location of the double bond nearest the terminal methyl group.

In order to assess model performance, diet estimates computed from the MLE method

will be compared to the true known diets. In Nordstrom et al. (2008), these calculations used a cumulative recorded diet intake for each individual over the assessment period between 35 and 75 days prior to each of the biopsy periods. This method ignores the presence of existing fat stores that the animals may have had when they arrived at the facility and assumes that there is complete or almost complete turnover of FAs in the selected period used for calculation. Shelley Lang has modified the calculations to attempt to account for the contribution of the blubber present at the time of admission to the overall blubber FA signature at each experimental time point. For each study animal, the cumulative mass intake of the prey fed was used to estimate the expected contribution of each prey species to the blubber fatty acid signature (as a proportion of diet) at Days 0, 21, and 42 of the experimental period. The body fat content of the study animals at the time of admission to the rescue facility was not known. Therefore, to account for the contribution of the blubber present at the time of admission to the overall blubber FA signature at each experimental time point, a baseline body fat content at admission was estimated, and then it was assumed that the FAs subsequently consumed were deposited with existing fat, and used as a single pool (see Iverson et al. (2004a)). Because the study animals were growing pups, it was also assumed that a fraction of the fatty acids consumed were immediately oxidized and not deposited. For all individuals, the amount of blubber present at the time of admission was estimated using the value for newborn harbour seal pups from Bowen et al. (1992) of 11.3% body fat. This value may be an overestimate for pups which may have lost significant body condition following the separation from their mother but it provided a reasonably conservative estimate in the absense of information on prior feeding history (i.e. duration of suckling before separation), the subsequent duration of separation, or body condition at the time of arrival to the facility (condition score or condition index; Lander et al. (2003)). Storage efficiency (the proportion of energy intake subsequently deposited) has not been examined for harbour seal pups, therefore, the value of 70% obtained for nursing pups in the closely related grey seal (*Halichoerus grypus*;, Mellish et al. (1999), Lang et al. (2011)) was used.

Before analysis, the FA signature of each seal was visualized using the bar plot shown in Figure 6.1. From this, we can see that 22:6n-3 and 20:5n-3 appear to be the most

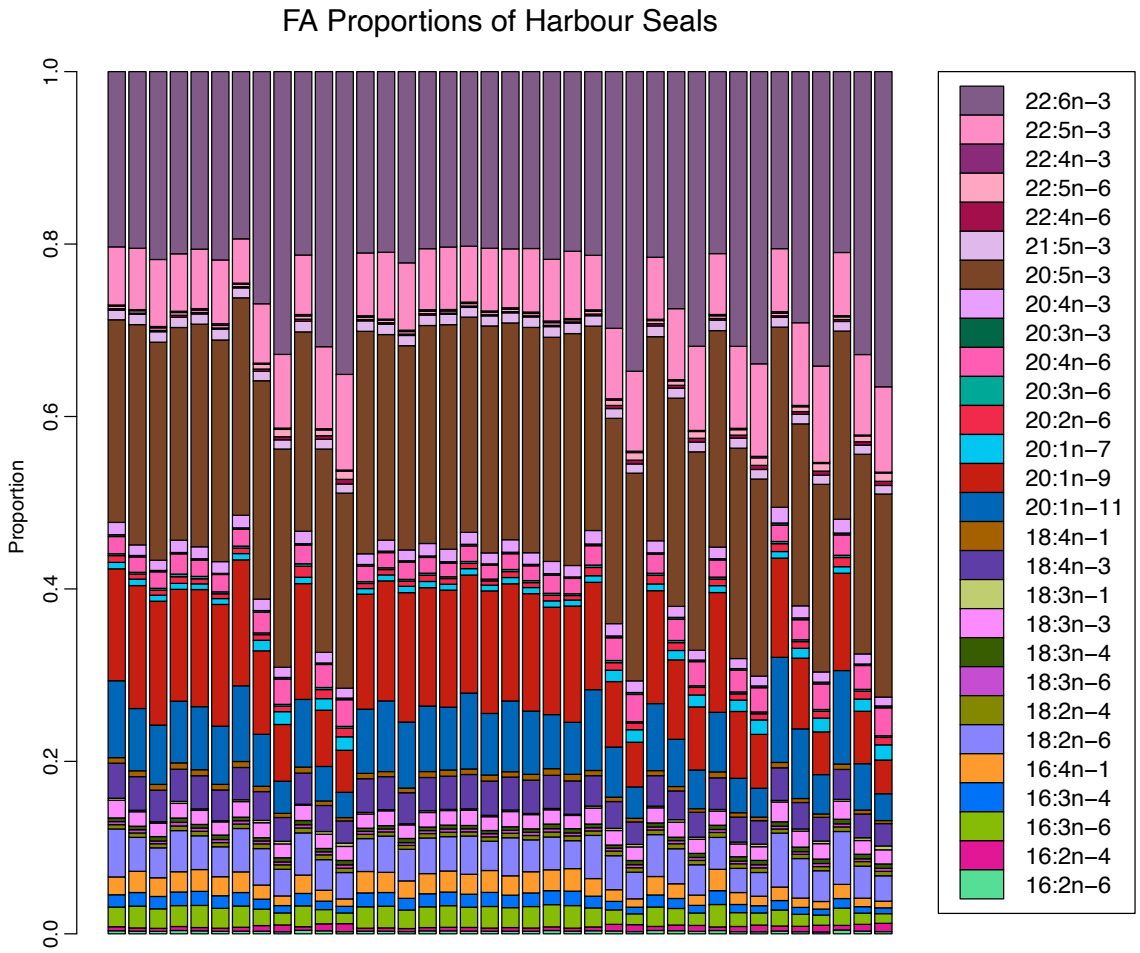

Figure 6.1: Proportion of FAs in each harbour seal sample, from all time points, with some replicates.

dominant FAs in the signatures, followed by 20:1n-9, 20:1n-11 and 22:5n-3. However, 20:5n-3 appears to be fairly consistent across individuals whereas 20:1n-9 and 20:1n-11 are varying larger amounts, relative to their proportion sizes. The less dominant FAs are difficult to assess in this plot, but may also contribute to differences among the diets.

6.2 Diet Estimates

The preybase used includes 11 species, one of which (surfsmelt) is broken into large and small subgroups. The species, their sample sizes and average lipid or fat content (%) is shown in Table 6.1. Fat content can be used to adjust the diet proportions

Species	n	Average Lipid (%)
Capelin	54	3.28
Coho	38	3.79
Eulachon	30	8.80
Herring	23	11.23
Mackerel	24	5.87
Pilchard	18	20.33
Pollock	17	6.88
Salmonoil	5	100.00
Sandlance	15	5.73
Squid	43	2.21
Sursfmelt lg	30	2.85
Surfsmelt sm	10	2.24

Table 6.1: Sample sizes and average lipid (%) of the 11 species (12 prey groups) included in the prey base for harbour seals.

so that they are not relative to the amount in the FA signature, but to the amount consumed. This allows us to take into account that species with higher fat content will contribute more to the FA signatures than those that are low in fat content. The true diets for the 38 harbour seal biopsies were obtained as described in Section 6.1, are shown in Table 6.2. As a visual description of Table 6.2, a ternary diagram is shown in Figure 6.2 that includes only the 3 species that are non-zero in the diets of every individual, namely herring, salmonoil and surfsmelt. We can see from the ternary diagram that the diets generally include only 2 of the 3 species in one individual, as most of the dots are very close to, if not on the edges. In the past, diets along the borders of the simplex have been trickier to deal with, particularly because of the presence of zeros.

Because analysis with all 11 species is computationally intensive, this analysis was only run once, without winsorizing. Winsorizing does not add to the computation time, however, as seen in Section 4.2, winsorizing does not appear to make a significant difference in the estimates. Since the diet of each individual seal was different, instead of plotting the diet estimates, the bias of each individual, that is, true diet - diet estimate, is plotted in a box plot shown in Figure 6.4. Note that the scale is not the same for all species in this figure. From this, we can see that for most

Seal	Capelin	Coho	Eulachon	Herring	Mackerel	Pilchard	Pollock	Salmonoil	Sandlance	Squid	Surfsmelt
1	0	0	0	0.88	0	0	0	0.07	0	0	0
2	0	0	0	0.94	0	0	0	0.04	0	0	0
3	0	0	0	0.95	0	0	0	0.03	0	0	0
4	0	0	0	0.89	0	0	0	0.05	0	0	0
5	0	0	0	0.95	0	0	0	0.02	0	0	0
6	0	0	0	0.96	0	0	0	0.02	0	0	0
7	0	0	0	0.84	0	0	0	0.09	0	0	0
8	0	0	0	0.31	0	0	0	0.03	0	0	0.63
9	0	0	0	0.18	0	0	0	0.02	0	0	0.79
10	0	0	0	0.81	0	0	0	0.08	0	0	0
11	0	0	0	0.29	0	0	0	0.03	0	0	0.64
12	0	0	0	0.16	0	0	0	0.02	0	0	0.80
13	0	0	0	0.92	0	0	0	0.04	0	0	0
14	0	0	0	0.94	0	0	0	0.03	0	0	0
15	0	0	0	0.95	0	0	0	0.02	0	0	0
16	0	0	0	0.90	0	0	0	0.05	0	0	0
17	0	0	0	0.94	0	0	0	0.03	0	0	0
18	0	0	0	0.87	0	0	0	0.06	0	0	0
19	0	0	0	0.94	0	0	0	0.03	0	0	0
20	0	0	0	0.92	0	0	0	0.04	0	0	0
21	0	0	0	0.95	0	0	0	0.02	0	0	0
22	0	0	0	0.88	0	0	0	0.05	0	0	0
23	0	0	0	0.95	0	0	0	0.02	0	0	0
24	0	0	0	0.77	0	0	0	0.11	0	0	0
25	0	0	0	0.27	0	0	0	0.04	0	0	0.65
26	0	0	0	0.15	0	0	0	0.02	0	0	0.81
27	0	0	0	0.84	0	0	0	0.08	0	0	0
28	0	0	0	0.32	0	0	0	0.03	0	0	0.62
29	0	0	0	0.19	0	0	0	0.02	0	0	0.78
30	0	0	0	0.81	0	0	0	0.06	0	0	0
31	0	0	0	0.24	0	0	0	0.02	0	0	0.71
32	0	0	0	0.13	0	0	0	0.01	0	0	0.84
33	0	0	0	0.62	0	0	0	0.25	0	0	0
34	0	0	0	0.16	0	0	0	0.04	0	0	0.74
35	0	0	0	0.09	0	0	0	0.02	0	0	0.86
36	0	0	0	0.65	0	0	0	0.15	0	0	0
37	0	0	0	0.15	0	0	0	0.04	0	0	0.76
38	0	0	0	0.08	0	0	0	0.02	0	0	87

Table 6.2: True diets of 38 captive grey seals at the Vancouver Aquarium.

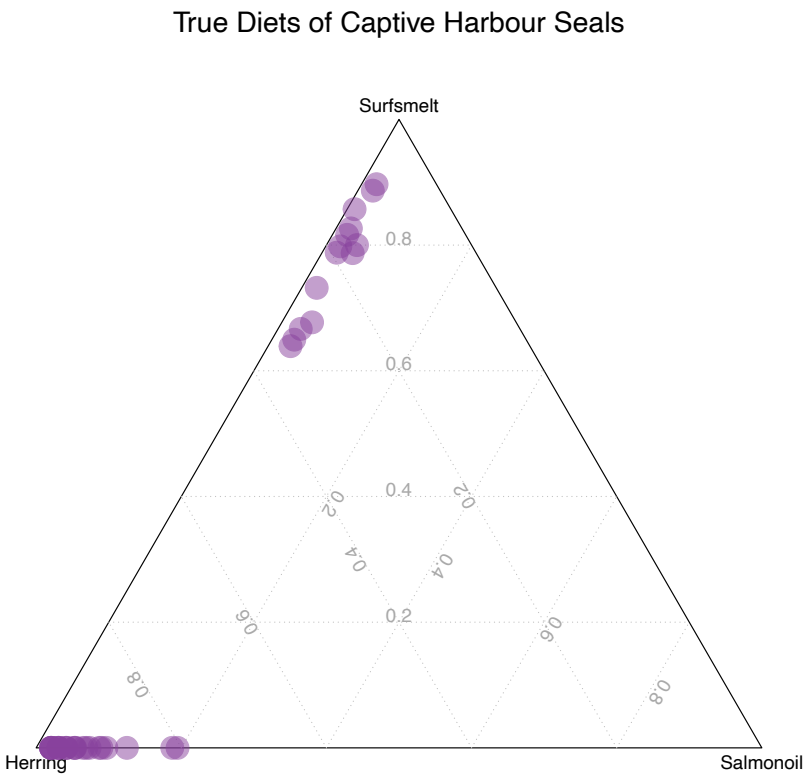

Figure 6.2: Ternary diagram of the true diets of 38 biopsies of harbour seals during a captive study at Vancouver Aquarium.

species that have true diet proportion 0, namely, capelin, coho, eulachon, mackerel and pollock, our estimate appears to be performing quite well, as the median bias is 0, with little to no variation. For other species with true proportion zero, namely pilchard, sandlance, and squid, both QFASA and MLE seem to have a more difficult time estimating the zero. Both QFASA and MLE seem to overestimate all three of these, pilchard by about 10%, sandlance by about 5% for MLE, 10% for QFASA and squid by around 5% for QFASA and 1% for MLE. All in all, even with zeros, our MLE method is performing as well, or better than QFASA. With the three species that are being consumed by the seals, herring is being slightly overestimated by both methods, with a large amount of variability in the estimates, salmonoil is being slightly overestimated, with small variability but with several large outliers for the MLE method, and survfsmelt has a median right on the 0 bias. However, there is an extremely large amount of variability in the estimates, with a tendency to overestimate the proportion. Again, our goal was to be comparable to QFASA in the estimation, which from this real life data set, seems to be the case, and in most situations, the MLE method estimated even better than QFASA.

The biases are relatively small, and thus the method appears to be estimating accurately. However, an explanation of the bias is still desired. To determine what causes the bias, FA signatures of the species are compared, first using a dendrogram shown in Figure 6.3. Prey species with similar FA signatures are difficult to differentiate from each other, and thus could explain over/under estimating. For example, pilchard and herring have very similar FA signatures, and we can see in Figure 6.4 that herring is underestimated by approximately 10% and pilchard is overestimated by approximately 10%. Therefore, the similarities in FA signatures may be creating a slight bias in estimation for these two species.

Next, we wanted to compare estimates using winsorized and non-winsorized preybases, but due to the length of computations with all 11 species, this was done using only the 3 species that have non-zero values in the true diets: herring, salmonoil and

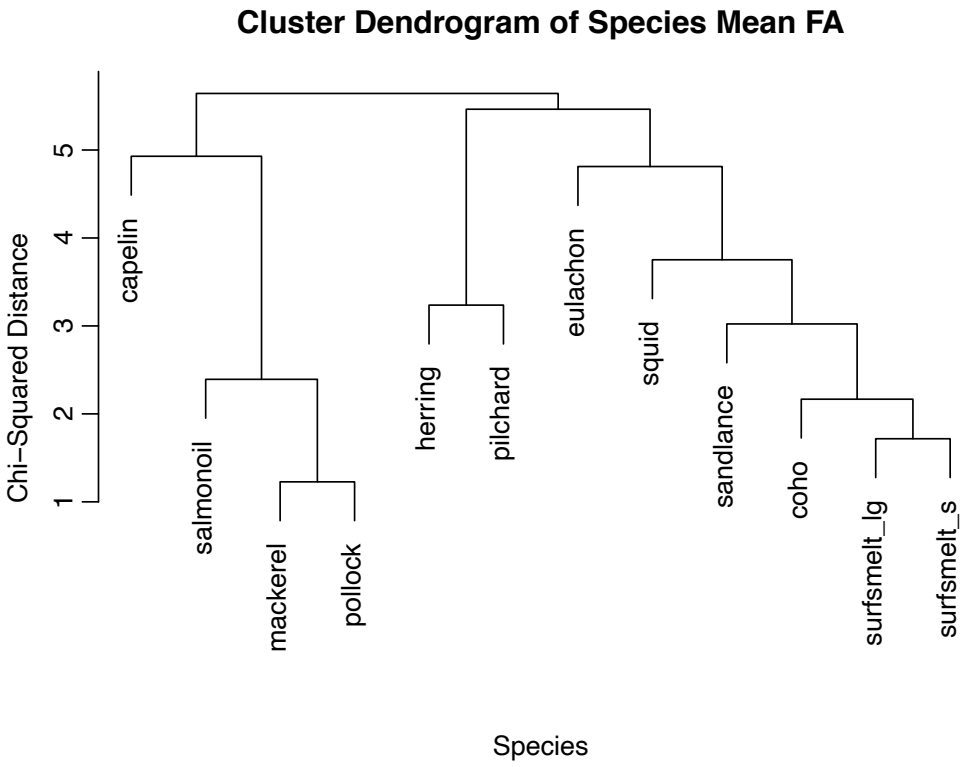

Figure 6.3: Dendrogram of 11 prey species included in the Nordstrom preybase, using the mean FA signatures, and chi-squared distances.

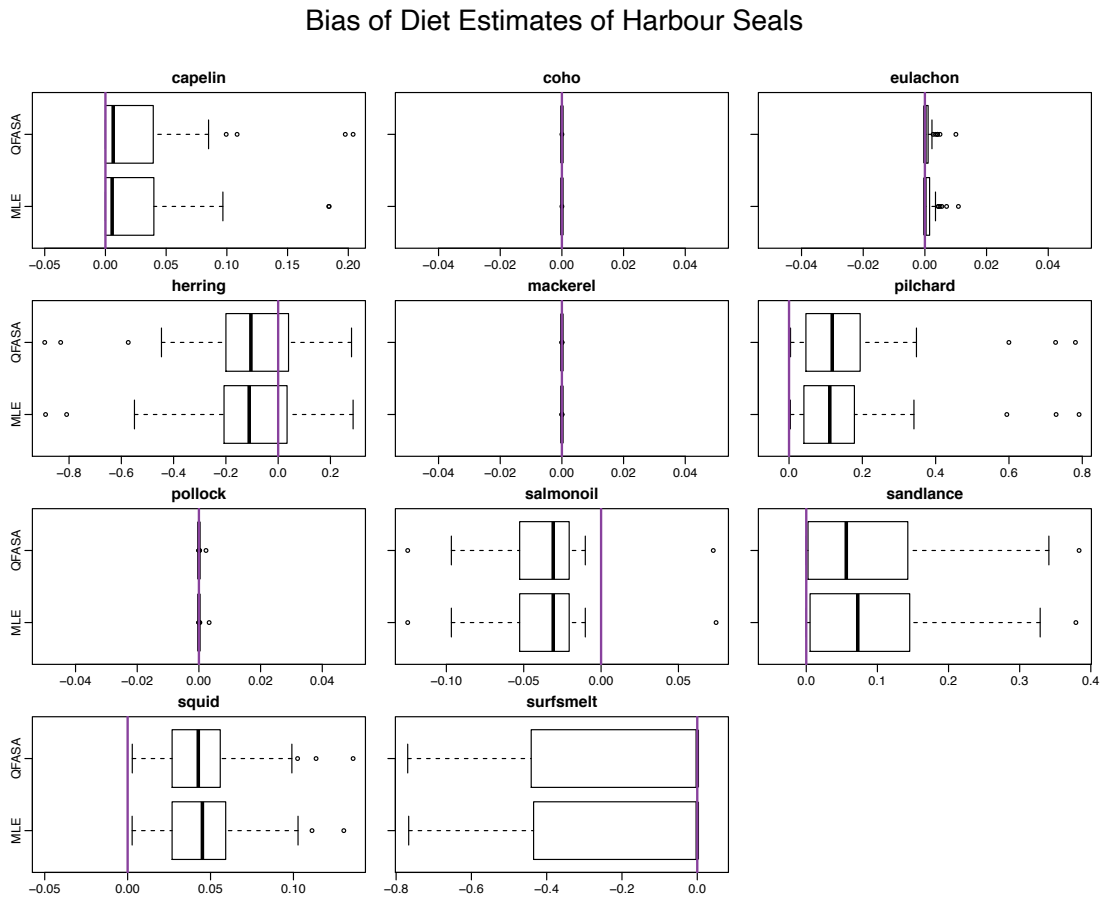

Figure 6.4: Bias (true diet - estimated diet) for 38 harbour seals, estimated using both MLE and QFASA.

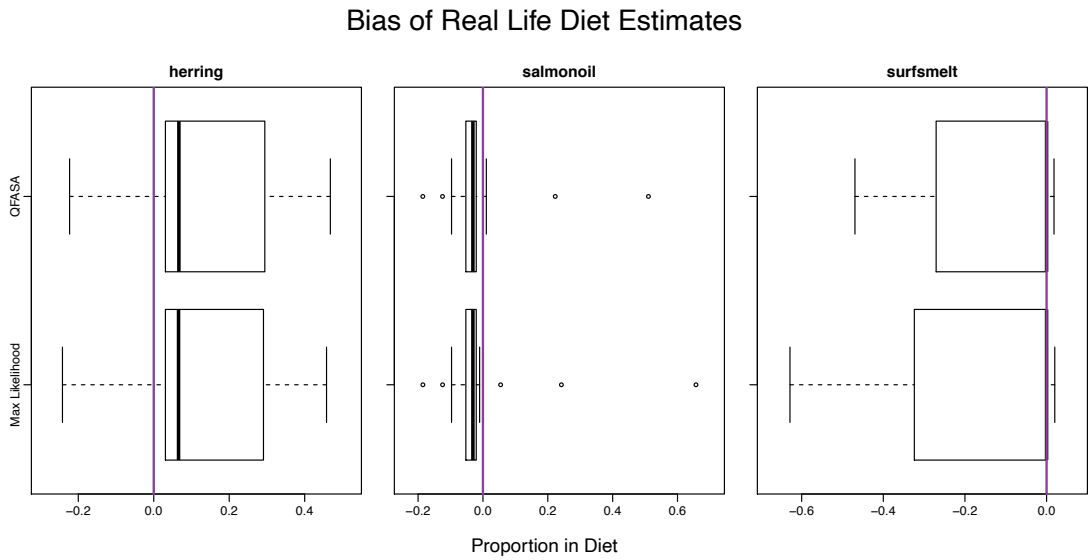

Figure 6.5: Bias of diet estimates for 38 real life seals, using only the species that were consumed, using the non-winsorized preybase.

surfsmelt. Similar to above, boxplots of bias of the estimates are shown in Figure 6.5 for the estimates with the non-winsorized preybase, and Figure 6.6 for the estimates with the winsorized preybase. Excluding the species with 0 diet proportions, we see similar results in the estimation of these diets: herring is being slightly overestimated with a wide variability, salmonoil is being slightly underestimated with a small variability, and surfsmelt is highly variable, tending towards underestimation. However, when we compare the winsorized and non-winsorized methods, there does not appear to be much difference, other than a slight decrease in the variability of the surfsmelt estimates. All in all, it does not appear that winsorizing has a significant effect on the diet estimates, nor does a lack of normality of the FA signatures, as despite the fact that most FAs tended to be non-normal, the MLE method is still performing well with the real life data.

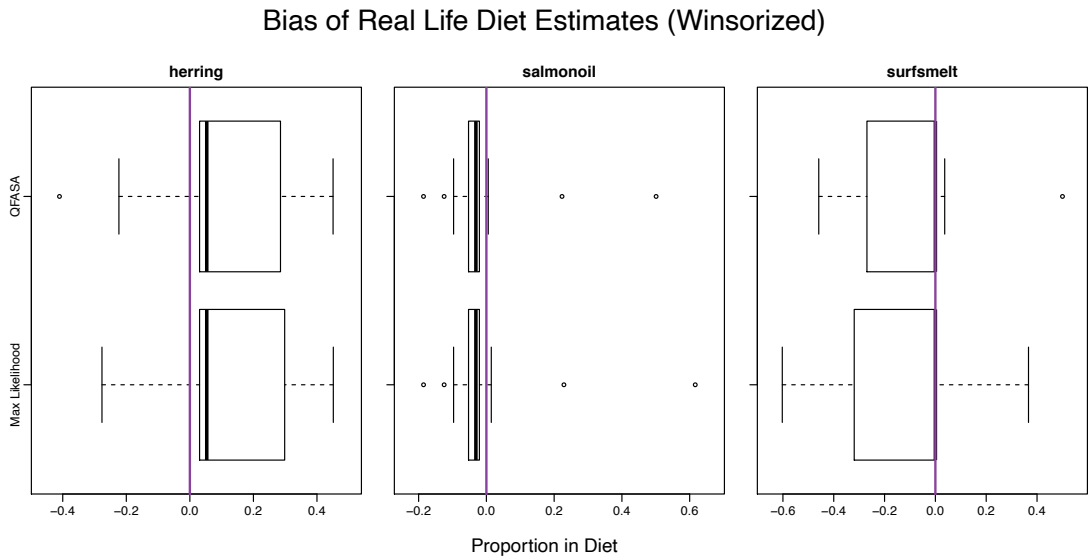

Figure 6.6: Bias of diet estimates for 38 real life seals, using only the species that were consumed, using the winsorized preybase.

6.3 Real Life Data with Covariates

6.3.1 Data Collection

Seal Samples

Full-depth blubber biopsies were collected between 1994 and 2015 from adult grey seals during the annual breeding season (December-January) on Sable Island, NS (43°55′ N, 60°00′ W) following the methods described in Beck et al. (2007). Blubber samples were collected from a total of 502 individuals (183 males, 319 females; Table 6.3) used in studies examining diet, energetics, foraging distribution and behaviour (Mellish et al. (1999), Lidgard et al. (2003), Noren et al. (2005), Austin et al. (2006), Breed et al. (2006), Beck et al. (2007), Lang et al. (2009), Lang et al. (2011), Lidgard et al. (2020)). Over the course of lactation, female grey seals do not mobilize blubber fatty acids in a uniform manner (Arriola et al. (2013)), therefore, all blubber samples used in this study were collected from lactating females prior to day 6 post post-partum. Biopsies were wrapped in aluminium foil and kept chilled for several hours until placed in a solution of chloroform containing 0.01% 2,6-di-tert-butyl-4-methyl-phenol (BHT) by weight and stored frozen until analysis. Lipids were extracted from

all blubber samples using a modified Folch method (see Budge et al. (2006)). FA methyl esters (FAME) were prepared from the extracted lipids using an acidic catalyst (the Hilditch method, see Budge et al. (2006)). FAME were analysed in duplicate using temperature-programmed gas liquid chromatography according to Iverson et al. (1997). Individual FA are reported as mass percent of total FA.

Prey Library

Fish and invertebrate samples were collected during stratified random bottom-trawl surveys conducted by the Canadian Department of Fisheries and Oceans (DFO) in the spring, summer, or fall on the Scotian Shelf (Northwest Atlantic Fisheries Organization (NAFO) Divisions 4V, 4W, and 4X), Georges Bank (5Z) and the Gulf of St Lawrence (4S and 4T) between 1990 and 2001 (see Budge et al. (2002)). Additional fish and invertebrate samples were obtained from research cruises and commercial fisheries in the Gulf of St Lawrence (4S and 4T) between 2002 and 2004. At collection, individuals or groups of individuals of each species were stored frozen at 20°C in airtight plastic bags until analysis.

Individual prey samples were thawed and fork length was measured to nearest 0.1 cm; body mass was measured to the nearest 0.1 g. Each individual was then homogenised in a food processor. To determine fat content, lipids were quantitatively recovered in duplicate from samples of the homogenised prey using a modified Folch method (Folch et al. (1957), Iverson et al. (2001)). FA methyl esters (FAME) were prepared and analysed using temperature-programmed gas liquid chromatography as described above. Individual FA are reported as mass percent of total FA.

From a total prey library of more than 3700 individual prey FA signatures (82 species), 1735 individual FA signatures were selected from 21 species of fish and invertebrates that were collected within the NAFO 4 Subarea only (excluding the Gulf of St Lawrence estuary). This selection area covers the main foraging range of the Sable Island grey seals. The 21 prey included in the selected library (Table 2) were those known to be eaten by grey seals based on previous stomach content and

faecal analyses (e.g. Bowen et al. (1993), Bowen & Harrison (1994)) or prey that were reasonably abundant and found at depths at which grey seals are known to forage (Beck et al. (2003b), Beck et al. (2003a)).

Following an exploratory analyses to determine if the FA signatures of the selected prey contained any hidden structure (see Bromaghin et al. (2017a)), some prey species within the set were subdivided into smaller clusters (Table 6.4). American plaice were separated into 2 clusters based on size (small, $\leq$25cm and large, >25cm). Substructure based on seasonal variation (collection month) was found in Atlantic butterfish, Atlantic herring, capelin and longhorn sculpin. These species were separated into clusters and only the clusters representing collection months in summer and fall were retained for the estimation of the diets. Pollock were separated into 2 clusters based on observed substructure, although the proximate cause for the substructure was unclear (there was no relationship to differences in size, season or collection location).

6.3.2 Results

First, we considered the simple case of one covariate: sex. Using the methodology from Section 5.1, $\boldsymbol{\beta}$ coefficients were estimated, yielding a summary diet for males and for females. These diet estimates, along with the mean (arithmetic) QFASA diet estimate is displayed in Figure 6.7. With such a large group, we are getting a summary diet of the 183 males and a summary diet of the 319 females for both MLE and QFASA methods. We can see that there seems to be more similarities within the estimates than there is within the sexes. MLE seems to yield near 0 estimates for all prey species except for redfish, of which it yields an estimate of nearly 1, for both males and females, whereas QFASA yields estimates of around 0.6 for both males and females for redfish, with several smaller estimates for pollock, capelin, cod and sandlance. It seems unlikely that these seals would be eating nearly all redfish, as this would be a restrictive diet to hunt for in the wild. Since this grouping yielded one summary diet for very large groups of predators, we explored what the summary diets would be by breaking the predators into smaller groupings.

Next, we added another covariate into our model. Year group was added into the

Year	Male	Female	Total
1994	20	26	46
1995	21	19	40
1996	4	32	36
1997	21	39	60
1998	7	8	15
1999	26	11	37
2000	16	26	42
2001	5	7	12
2002	4	12	16
2003	7	13	20
2004	11	22	33
2005	13	28	41
2006	0	10	10
2009	4	6	10
2010	6	8	14
2011	5	24	29
2012	0	14	14
2013	5	4	9
2014	4	6	10
2015	4	4	8
Total	183	319	502

Table 6.3: Sample sizes of male and female adult grey seals organized by collection year.

CommonName	ScientificName	Subgroup	n
American plaice	*Hippoglossoides platessoides*	small <25 cm large >25 cm	67 67
Atlantic butterfish	*Peprilus triacanthus*		26
Atlantic cod	*Gadus morhua*		109
Atlantic herring	*Clupea harengus*	July-September	121
Atlantic mackerel	*Scomber scombrus*		32
Capelin	*Mallotus villosus*	July September	27 21
Longhorn sculpin	*Myoxocephalus octodecemspinosus*	September	25
Northern sandlance	*Ammodytes dubius*		148
Northern shortfin squid	*Illex illecebrosus*		35
Pollock	*Pollachius virens*	Group 1 Group 2	35 18
Redfish	*Sebastes sp.*		54
Sea raven	*Hemitripterus americanus*		71
Silver hake	*Merluccius bilinearis*		58
Smooth skate	*Malacoraja senta*		33
Snake blenny	*Lumpenus lumpretaeformis*		18
Thorny skate	*Amblyraja radiata*		83
White hake	*Urophycis tenuis*		80
Winter flounder	*Pseudopleuronectes americanus*		50
Winter skate	*Leucoraja ocellata*		40
Witch flounder	*Glyptocephalus cynoglossus*		24
Yellowtail flounder	*Limanda ferruginea*		156
			1398

Table 6.4: Sample sizes of prey species and their subgroups used in the FA analysis.

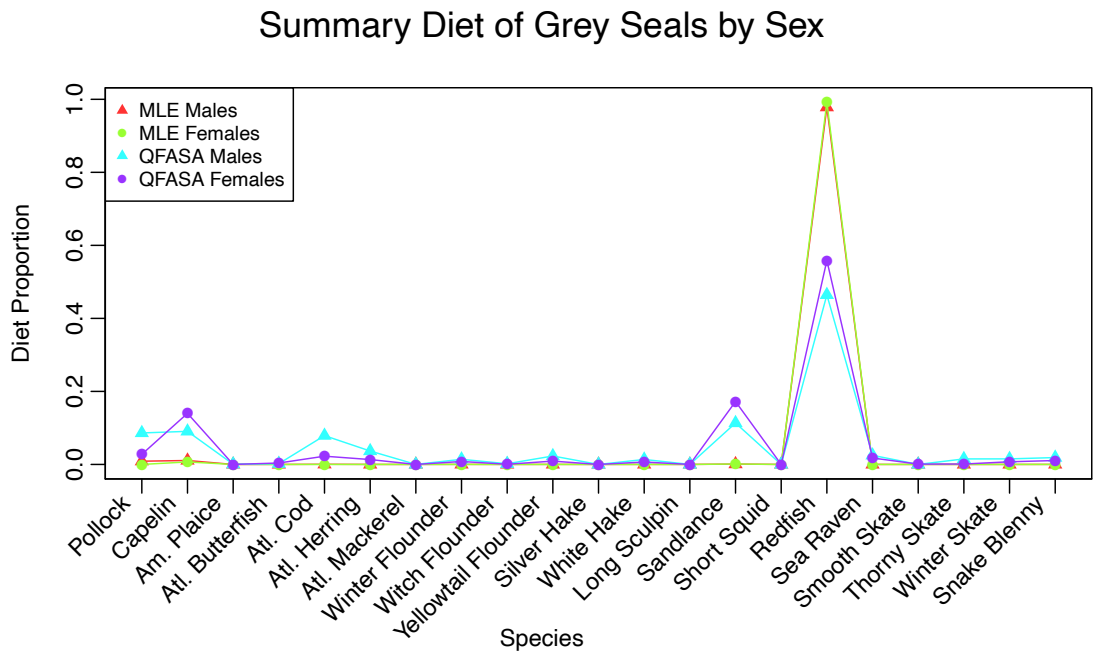

Figure 6.7: Summary diet estimates pbtained using MLE method and the average QFASA diet estimates grouped by sex of grey seals.

model, which takes on three possible values, 1994-1997, 1998-2004, and 2004-2015. They refer to different periods with different population growth rates for the Sable herd: exponential growth, a period during which population growth slowed (from exponential to a new rate), and then a period of stable population growth rate at the new level. Conveniently they divide into roughly equal group sizes in terms of numbers of samples in each. This covariate is added into the model using two indicator variables. Once again, we get a summary diet estimate for each unique set of covariates. Despite now having 6 unique groups (Male with the 3 growth periods, and female with the 3 growth periods), we obtained similar results for the diet estimates, as shown in Figure 6.8. Once again, redfish seems to dominate the diets for both males and females, using either ML or QFASA methods. However, once again, QFASA depicts a slightly more diverse diet, having only 30-50% on redfish, 10-20% on sandlance, and small percentages on pollock, capelin, Atlantic cod and yellowtail flounder, for both males and females. For females, ML estimates nearly 100% on redfish for all time periods, whereas for males, we see approximately 60% redfish with the remainder on pollock for the period of exponential growth, 80% redfish and 20% pollock for the period of decreasing rate, and nearly 100% redfish and a small percentage of capelin during stable growth rate.

6.3.3 Inference on Real Life

Similar to the simulation study, after adding covariates into the model, the next step is to determine if there is a statistically significant difference among the groups created by the covariate. Thus, we want to first see if there is a difference in diet between male and female grey seals, and also if there is a difference among the year groups. However, with 502 grey seals and 24 species groups, the bootstraps required for inference are computationally intensive, taking lots of time and memory. Therefore, to reduce computation, only the 5 prey species groups that had non zero diet estimates from Section 6.3.2 were included: capelin (July and September), redfish, Northern sandlance, and Atlantic cod. Then, as in Section 5.4, bootstraps were performed by generating pseudo-predators with true diets obtained from the ML $\boldsymbol{\beta}$ estimates. Then the ML method is used to get bootstrap estimate, $\boldsymbol{\beta}_r$. This is then repeated $r = 50$

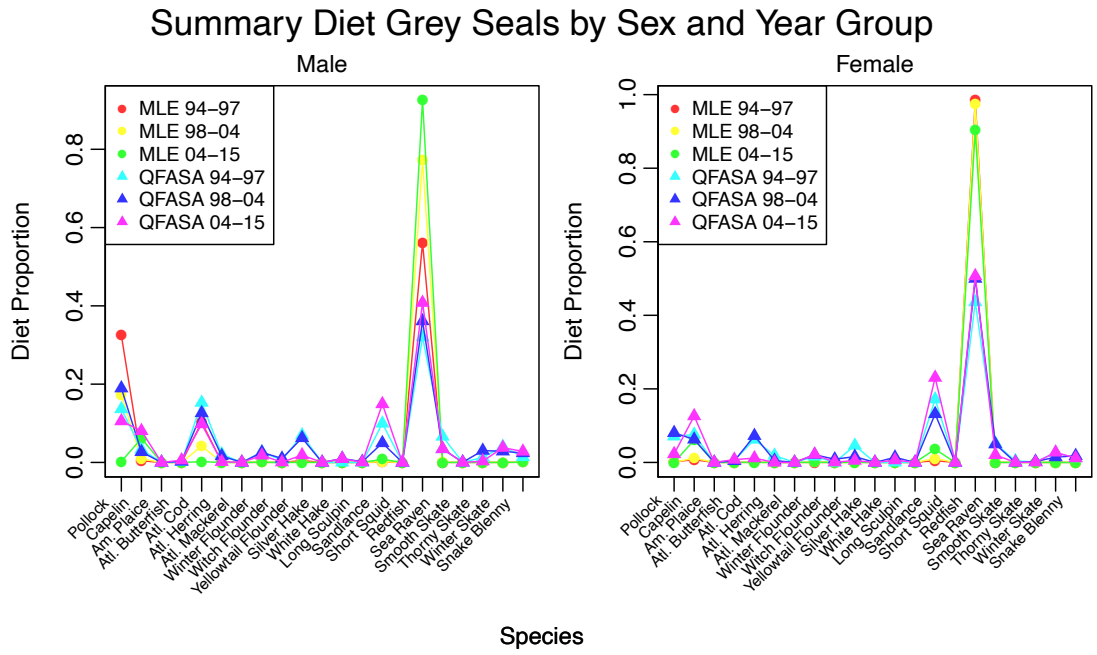

Figure 6.8: Summary diet estimates obtained using MLE method and the average QFASA diet estimates grouped by sex and year group of grey seals.

times.

To explore the difference in diet estimates between the full and reduced prey sets, summary diet estimates for males and females, from both MLE and QFASA are displayed in Figure 6.9 using the full prey set (with the 4 prey extracted and closed) and the reduced set. Although the proportions on certain species (for example, Atlantic cod) are different when using the reduced set compared to the full set of prey, the estimates are closer to those of QFASA when using the reduced set. As QFASA has been a widely used, tested and approved method in the past, it tells us that our reduced prey set estimates may be more reliable than with the full set. Also, it is generally known among biologists that there is a difference in diet between male and female grey seals, but with the full prey set, they are nearly identical. This could be indicating an identifiability problem with the full 24 prey species. For these reasons, and because of the computational complexity of the model with large prey sets, we will trust and use the estimates from the reduced prey set for the bootstrapping.

Based on 36 bootstraps (50 bootstraps, but those that did not converge were removed), using sex as a covariate, we obtained observed and critical distances of 19.92 and 3.08 respectively. As the observed distance is larger than the critical distance, we reject H_0 for H_a and conclude that at 5% significance, there is a difference in true diet between male and female grey seals. This validates our methods, as it is known among biologists that these diets differ. Therefore, as a preliminary study, this shows great promise that this inference technique is a valid way to test for differences in diet among groups.

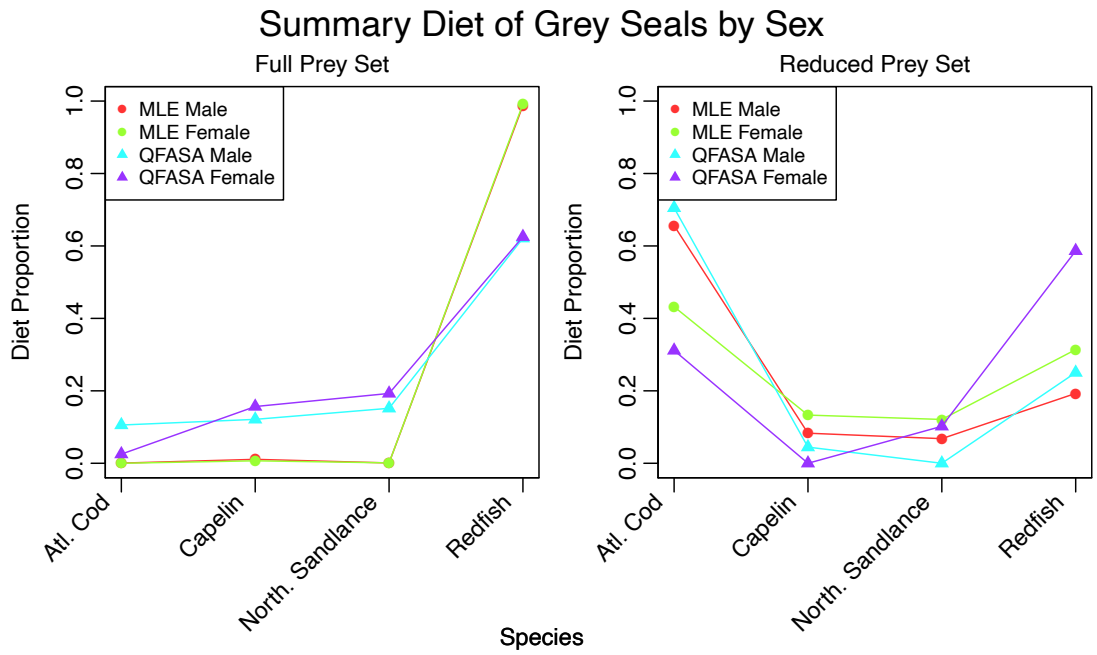

Figure 6.9: Summary diet estimates obtained using MLE method and the average QFASA diet estimates of grey seals by sex, using full (reclosed) and reduced prey sets.

Chapter 7

Conclusions

In this final chapter, the results, and recommendations discussed throughout the preceding chapters are summarized, as well as future research. It has been divided into ML diet estimation, inference with covariates, and real life analysis.

7.1 Summary

Our research was motivated by the need for an improvement in the statistical models used in quantitative fatty acid signature analysis (QFASA) proposed in Iverson et al. (2004a). The original method used given samples of predator and prey FA signatures from assumed prey species in the diet, and the estimated diet proportions were the weights that minimized the distance between the predator FA signatures and a weighted mixture of mean prey FA signatures for each prey species. This model did not include any variability in the predator or prey FA signatures, nor did it allow a way to include covariates such as age, sex, and location into the estimation process. Thus, we explored several improvements on this methodology, including random effects for unobserved prey FA signatures, a maximum likelihood approach to the model, and covariates in the likelihood function, allowing for inference in the same step as estimation.

In this thesis, we explored this maximum likelihood methodology using both simulations and two real life data sets. The methods and corresponding results are summarized below.

7.1.1 Maximum Likelihood Diet Estimation

Developing a maximum likelihood estimation method for diet proportions of predators based on QFASA required careful consideration of what is happening both biologically, and statistically. Since both the FA data and the diet estimates themselves are

compositional in nature, we began in Chapter 2 by reviewing and summarizing the existing techniques for analysing compositional data. In Chapter 3, we explored a maximum likelihood approach to QFASA. The first improvement we considered was based on the fact that the prey FA signatures collected in the samples are not from the exact individual prey consumed by the predators. Therefore, we included the prey FA signatures in our model as unobserved random effects. The sampled prey allowed us to use empirical summary statistics as estimates for parameters of the random effects' distributions.

Because our data is compositional in nature, transformations are required in order to perform multivariate analyses. Therefore, the weighted mixture of diet proportions and prey FA signatures could be considered in two spaces: the simplex, where we would take a mixture of untransformed FA signatures, or real Euclidean space, where we would take a mixture of transformed FA signatures. After considering the biological implications of both techniques, we decided that taking the mixture on the untransformed scale made the most sense, as the FAs would be absorbed into the predator tissues without scaling. This created a major challenge as the distribution of a mixture of compositions is not trivial. Thus, approximations based on Aitchison & Bacon-Shone (1999) were required to approximate the distribution of the convex linear combination of diet proportions and prey FA signatures. These approximations also had to be adjusted to incorporate a more recent and recently accepted and popular transformation, the isometric log-ratio transformation.

With the random effects and the unknown mean and variance-covariance matrices of the prey FA signatures, the number of parameters was much higher than the sample sizes. This created an identifiability problem. In order to lower the number of parameters being optimized, empirical means and variance-covariance matrices were computed for the prey species, and a diagonal matrix was assumed for the variance-covariance matrix for the error perturbations. This significantly decreased the computational difficulties and allowed the model to be run with relatively small sample sizes.

In Chapter 4, simulation studies were developed and run to determine how the ML

method behaved in practice. First, prey FA signatures were generated both parametrically, by randomly sampling from the multivariate normal distribution with mean and variance-covariance matrices estimated using sampled prey, and non-parametrically, by randomly sampling a prey FA signature from each species in the sampled prey base. Then, using 20 true diets, equally spaced throughout the simplex, pseudo-predators were generated by taking a convex linear combination of the true diets with the prey FA signature from each prey species. For both techniques, an error perturbation is randomly sampled from the multivariate normal distribution with mean vector $\mathbf{0}$, and variance-covariance matrix a diagonal matrix with 0.001 on the diagonal. This is then back transformed onto the simplex using the inverse ilr transformation, and perturbed with the generated FA signature. The resulting composition is the pseudo-predator FA signature.

Using these true diets, and three groups of 4 prey species, the behaviour of the maximum likelihood method was explored across a variety of situations. In most cases (other than a few non-parametric simulations), our ML method yielded estimates comparable, if not more precise and accurate, to traditional QFASA estimates. While the ML method is slower to run, it allows for important improvements over the original QFASA method, including but not limited to, inference and the inclusion of covariates.

Bootstrapped confidence intervals of the diet estimates were obtained by using marginal percentiles of diet proportion replicates. These bootstrap replicates were obtained by using the ML estimates as the true diet, and generating r pseudo-predators. The resulting diet estimates from optimizing the likelihood with the pseudo-predators are the bootstrap replicates.

The simulation results for the bootstrap CIs show that while our method obtains accurate and precise diet proportions, true diets on the edges of the simplex are less likely to be contained in the marginal CIs. This is due to the fact that the diet proportions are restricted to be between 0 and 1, and thus we will never obtain estimates below 0 or above 1. Therefore, in the best case, we will obtain a percentile on the

true value, but generally will be slightly above (for 0) or slightly below (for 1) the true values. A larger study of the coverage probabilities is an area for future work, to better explore the behaviour of these CIs.

Covariates

Covariates were included into the ML method described above by way of a link function. This link function is the same as that used in Dirichlet regression, and uses unknown, unbounded regression coefficients to obtain compositional diet proportions, that are constrained to being positive, between 0 and 1, and summing to 1. This link function is plugged into the likelihood in place of the diet proportions, and the likelihood is then optimized over the regression coefficients. When covariates are included in the model, a summary diet estimate is obtained for each unique set of covariates. For the simulations, we first started with one simple indicator variable covariate, and then to one covariate represented by two indicator variables.

For our simulations, the simple case began with assuming two groups, such as male and female. This is represented by one indicator variable. Thus, the results will yield a summary diet for the males, and a summary diet for the females. To see the performance of our new method with covariates, we compared the summary diet estimates to the empirical mean diet estimates obtained from QFASA. Similar to the simulations without covariates, equally spaced diets were used to ensure that the method is working throughout the simplex. All possible combinations of 2 of 10 such diets were used, and in all instances, the summary diet estimates were comparable, if not better than the mean QFASA estimates, displaying once again the accuracy and utility of our model.

Inference

One of the goals of this ML model was to test for differences among groups. Previously, this was done by getting diet estimates and then modelling the estimates with

covariates. This two step process loses information about the precision of the estimates when modelling with the covariates. So, we devised a way to perform inference in the same stage as estimation. After estimating the ML regression coefficients, $\boldsymbol{\beta}$, bootstrap replicates are obtained in a similar fashion as with the confidence intervals. Taking the ML estimates of $\boldsymbol{\beta}$, the corresponding diet estimates are obtained from the link function. These diet estimates are then used as the true diets to generate n pseudo-predators. These pseudo-predators are then run through the ML model to obtain bootstrapped regression coefficient replicates, $\boldsymbol{\beta}^r$. This is repeated r times.

Using these bootstrap replicates, we can test if the true diets from all unique groups are different by testing if all the regression coefficients (apart from the intercepts) are different from 0. We do this using a method proposed in Olive (2016) which uses the mean and variance-covariance matrix of the bootstrap replicated coefficients to obtain a Mahalanobis distance between the replicate and the mean coefficients for every replicate. Then, the $(1-\alpha)^{th}$ percentile of these distances is compared to the the Malahanobis distance between the mean replicate coefficient and 0. If this last distance is larger than the percentile distance, we will reject H_0 for H_a and conclude that there is a difference among the diets. Otherwise, we will fail to reject H_0.

To determine if the inference method is behaving as expected, first, all possible combinations of 2 of the 10 diets used in the covariate simulations were compared, and three were chosen to represent no effect, a small effect size, and a larger effect size based on the chi-squared distance between diets. For each effect size, 10 pseudo-predators for each group (20 in total) were generated, and ML estimates of the regression coefficients are obtained. Then, using the method described above, the hypothesis test is performed to determine if the groups have significantly different diet proportions. For all groups, using $\alpha = 0.01$, $\alpha = 0.05$, and $\alpha = 0.10$ as significance levels, the correct decision was made. That is, when there was no effect, the null hypothesis was not rejected, and for both small and large effects, the null hypothesis was rejected. While there was not enough time here to obtain power simulations, this shows great promise for this technique for detecting differences in diets among different groups.

Next, to add a level of complexity, three groups were assumed, such as pup, adolescent and adult, represented by 2 indicator variables. Diets for the three groups were chosen based on the mean pairwise chi-squared distances between the three diets, to represent no effect size (all three diets are equal), small single effect size (two diets equal, one slightly different), large single effect size (two diets equal, one largely different), small effect size (all three diets slightly different), and large effect size (all three diets largely different). For each effect size, 10 pseudo-predators from each group (30 total) were generated using the true diets, and the ML estimates of regression coefficients were obtained. Using these as the true coefficients, 50 bootstrap replicates were obtained, and the hypothesis testing method described above is used to determine if there is a difference in diets between the groups. Once again, for all groups, using significance levels $\alpha = 0.01$, $\alpha = 0.05$, and $\alpha = 0.10$, the correct decision was made every time. That is, when there was no effect, we failed to reject H_0 for H_a, and in all other cases, we rejected H_0 for H_a, concluding that there is a significant difference between the groups.

Real Life Data

For analyses on real life data sets, two different sets were used. The first is a captive feeding study on harbour seals conducted at the Vancouver Aquarium by Nordstrom et al. (2008). This study has approximate known diets, therefore we were able to compare our estimated diets from both QFASA and our ML method to the known parameters. First, we used all prey species, and estimated the diet proportions using the two methods. In every case, ML estimates were comparable to QFASA estimates. Since our method has statistical benefits over QFASA (inclusion of variability as well as inference), comparable estimates to QFASA is all we were hoping for. To make estimation a bit easier, we also reduced the number of prey species included to only the 3 species which had non zero true diet proportions. Once again, the estimates are comparable, both in median value, and in spread, to that of QFASA. This was done using raw FA signatures as well as winsorized FA signatures to improve the normality of the data. There was no significant improvement with winsorizing, but it was used throughout the rest of the real life analyses, as it does not require extra time, yet

improves the validity of the normality assumption.

The second real life data set contains FA signatures from adult grey seals on Sable Island, NS. For the 502 individuals, the sex, the age, as well as the year group (which describes the level of population growth on Sable island at the time; exponential growth, slowed exponential growth, and or stable growth rate). Sex and year group are both categorical variables and thus are represented by indicator variables in our model. First, we just looked at sex, obtaining a summary diet for males, and a summary diet for females. We compared this with the mean QFASA diet estimates for all males and that for females. While most of the 0 proportions were estimated identically for both methods, the MLEs for both males and females show nearly 100% on redfish. The QFASA estimates show between 50 and 60% redfish, with small proportions also on sandlance, Atlantic cod, capelin and pollock.

Lastly, inference was done on the real life set, using sex as a covariate. The observed and critical distances obtained were 19.92 and 3.08 respectively, allowing us to reject H_0 for H_a and conclude that there is a difference in diets between male and female grey seals. This has been widely assumed among biologists for some time, and therefore we can say our method made the correct decision. This demonstrates once again the validity of using this inference technique to test for differences in diet.

7.1.2 Future Research

This thesis provides a base for the ML methodology and demonstrates the validity and usefulness of such a model. However, there are still many areas that can be improved or added to. The first of which is the efficiency of the code. While we have established that the code I have created behaves properly, it is quite slow and laborious to run. Making this code more efficient would make computations faster, which in turn would make larger simulation and real life studies more realistic.

Secondly, several new techniques for QFASA were mentioned throughout this thesis, namely the augmented matrix approach proposed in Bromaghin et al. (2016) and the simultaneous estimation of calibration coefficients proposed in Bromaghin et al.

(2017b). Implementing these into this novel ML method and performing further testing would be beneficial to see if these techniques improve our methods further. Currently, a graduate student under Connie Stewart is working on the calibration coefficient problem which should prove useful to future work in this area.

Appendix A

Code

A.1 MLE Method

A.1.1 R Code

```
## Using full FAset
FAset <- read.csv("/misc/home/steevesh/Comparison.Study
/FAset.csv",header=T,sep=",")
FAset <- as.vector(unlist(FAset))
## Remove this FA for Grey Seals
FAset <- FAset[!(FAset=="c16.4w3")]

# Calibration coefficients to use
cal.mat.orig <- read.csv("Cal.Mat.csv", header=T, sep=",")
cal.mat <- cal.mat.orig$CC
names(cal.mat) <- cal.mat.orig$FA

# Extracting FAset
cal.mat <- cal.mat[FAset]

# Prey database
preybase <- read.csv("/misc/home/steevesh/Comparison.Study
/Prey.csv",header=T,sep=",")

# prey species
spec <- unique(preybase$Species)
# Order the species so nothing gets mixed up in the estimation process
spec <- spec[order(spec)]
```

```
I <- length(spec)


# Extracting only the prey species to be including,
# and making sure its ordered correctly
for(j in 1:length(spec)){
  if(j==1) {preybase.spec <- preybase[preybase[,2]==spec[j],] }
  else{
    preybase.spec <- rbind(preybase.spec,
     preybase[preybase[,2]==spec[j],])
  }
}
preybase <- preybase.spec
preybase[,2] <- droplevels(preybase[,2])
# Making sure it sums to 1
preybase[,-(1:3)] <- preybase[,-(1:3)]/apply(preybase[,-(1:3)],1,sum)

 # Winsorize the data set
preybase.w <- matrix(NA,nrow=nrow(preybase), ncol=(ncol(preybase)-3))
rownames(preybase.w) <- preybase[,c("Species")]

for(i in 1:I){
  preyi <- preybase[preybase[,c("Species")]==spec[i],][,-(1:3)]
  for(j in 1:(ncol(preybase)-3)){
    iqr <- IQR(preyi[,j])
    vals <- quantile(preyi[,j], c(0.25, 0.5, 0.75))
    low <- vals[1] - 1.5*iqr
    high <- vals[3] + 1.5*iqr
    preybase.w[rownames(preybase.w)==spec[i],j] <-
      Winsorize(preyi[,j], minval = low, maxval=high)
```

```
  }
}
preybase.w <- preybase.w/apply(preybase.w,1,sum)

colnames(preybase.w) <- colnames(preybase[,-(1:3)])

preybase.q <- preybase.w

# Extract FAs from prey data base
preybase.q <- preybase.q[,FAset]
preybase.q <- preybase.q/apply(preybase.q,1,sum)
sort.preytype <- order(rownames(preybase.q))
preybase.q <- preybase.q[sort.preytype,]



# Transforming full prey base
preybase.t <- mod.zeros.FA.sig.mat(preybase.q,0.00005)
#analytic precision is two decimals after the percentage.
rownames(preybase.t) <- rownames(preybase.q)
preybase.t <-ilr(preybase.t)
preybase.w <- matrix(NA, nrow=nrow(preybase.t), ncol=ncol(preybase.t))
rownames(preybase.w) <- preybase.q[,"Species"]

# getting the number sampled from each species
n <- tapply(preybase$Species, preybase$Species,
length)[unique(preybase$Species)]



## Estimating the means and variance covariance matrices of
#our prey data
prey.var <- vector("list", I)
prey.mt <- matrix(0, nrow=I, ncol=length(FAset)-1)
```

```
prey.m <- matrix(0, nrow=I, ncol=length(FAset))

for(i in 1:I){
  preyi <- preybase.q[rownames(preybase.t)==spec[i],]
  preyit <- preybase.t[rownames(preybase.t)==spec[i],]
  prey.var[[i]] <- variation.acomp(preyi)
  prey.mt[i,] <- apply(preyit,2,mean)
  prey.m[i,] <- apply(preyi,2,mean)
}

D <- length(FAset)
V <- ilrBase(D=length(FAset))
G <- V%*%t(V)

## Calculating the pooled variance estimate
Spool <- matrix(0, D-1,D-1)
for(j in 1:I){
  Sigma <- -0.5*t(V)%*%G%*%prey.var[[j]]%*%G%*%V
  Spool <- Spool + (n[[j]]-1)*Sigma
}
Spool <- Spool/(sum(n) - length(n))

# Load in predator Data
Seals <- read.csv("CNordstrom_seals.csv")

# remove the columns not corresponding to FAs
 Sealsj <- Seals[,-(1:4)]
Sealsj <- Sealsj[,FAset]

### Getting Start Values
n.pred <- nrow(Sealsj)
```

```
# Start diet estimates from QFASA
test <- p.QFASA(predator.mat=Sealsj,prey.mat=prey.m,
cal.mat=cal.mat, dist.meas=2, start.val=rep(1/I,I),
 ext.fa=FAset )

# Remove the last column of diet estimates, as for our
#method we do not use it
start.val <- test$`Diet Estimates`[,-I]

# Seals after calibration coefficients
Sealsc <- matrix(NA,nrow=nrow(Sealsj), ncol=ncol(Sealsj))
for(i in 1:length(test$`Additional Measures`)){
  Sealsc[i,] <- test$`Additional Measures`[[i]]$ModFAS
}

# Modify seal data after calibration coefficients in case there are 0s
pred <- mod.zeros.FA.sig.mat(Sealsc,0.00005)
# transform predator data
ilr.seals <- ilr(pred)

#estimating errors
ers <- matrix(NA, nrow = 50*n.pred, ncol = ncol(Sealsj))

for(j in 1:50){
# generate seals from diet estimates without error included
  yest <- matrix(NA, nrow = n.pred, ncol = ncol(Sealsj))
  for(i in 1:nrow(yest)){
    yest[i,] <- pseudo.seal.norm(prey.mt, Spool,
     test$`Diet Estimates`[i,])
  }
  Sealsj <- acomp(Sealsj)
```

```
  yest <- acomp(yest) #without error, untransformed
  lb <- (j-1)*n.pred + 1
  ub <- j*n.pred
# error is the real life FA signatures with error minus the
# generated signatures without error
  ers[lb:ub,] = -yest + Sealsj
}
print(ers)
ers <- acomp(ers)
# estimated diagonal of variance covariance matrix of the predator
sep.start <- diag(-1/2*t(V)%*%G%*%variation(ers)%*%G%*%V)

# splitting the diagonal into quantiles to estimate to minimize
 number of unknown parameters
quan <- quantile(sep.start)
quan.start <- numeric(4)
groupind <- numeric(length(sep.start))
for(j in 1:4){
  quan.start[j] <- mean(sep.start[sep.start>=quan[j] &
    sep.start<=quan[j+1]])
  groupind[sep.start>=quan[j] & sep.start<=quan[j+1]] <- j
}

#### Running the program on the Real Life Seals

### create and load function from .cpp template
compile("ErrorModelSimpleEquant.cpp", flags="-Wno-unused-variable")
# create .o and .so
#dyn.unload(dynlib("ErrorModelSimpleEquant"))
dyn.load(dynlib("ErrorModelSimpleEquant")) # load .so

# Parameter start values
```

```
parameters <- list(alpha = start.val,
                   z  = array(rep(prey.mt), c(nrow(prey.mt),
                     ncol(prey.mt), n.pred)),
                   sepsilon = quan.start
)

#Data to send to tmb
data <- list(y = ilr.seals,
             #            x = preybase.t,
             n=n,
             varz=Spool,
             mu = prey.mt,
             V=ilrBase(D=(ncol(ilr.seals)+1)),
             sind=groupind
)

objnt <- MakeADFun(data,parameters,random="z")
npars <- length(objnt$par)

reports <- objnt$report()
reports <- objnt$report()
reports

# constraining alphas to be between 0 and 1
# Full parameter list
lb <- rep(0,npars)
ub <- c(rep(1,(npars-length(quan.start))), rep(Inf,length(quan.start)))

# alphas are fixed
#lb <- rep(0,npars)
#ub <- rep(Inf,npars)
```

```
# sepsilon is fixed
# lb <- rep(0,npars)
# ub <- rep(1,npars)

# sepsilon is fixed
#al.sum <- function(pars){
#  alpha <- matrix(pars,ncol=4)
#  return(apply(alpha,1,sum))
#}
#

# Full parameter set
al.sum <- function(pars){
  npars <- length(pars)
  alpha <- matrix(pars[1:(npars-length(quan.start))], ncol=I-1)
  return(apply(alpha,1,sum))
}

# Full parameter set
optnt <- solnp(pars=objnt$par, fun=objnt$fn, ineqfun=al.sum,
ineqLB = rep(0,n.pred), ineqUB = rep(1,n.pred), LB=lb,
UB=ub, control=list(delta=0.0001, tol=0.00001))

L <- optnt$values

alpha <- matrix(optnt$pars[1:((I-1)*(n.pred))],nrow=n.pred)
seps <- optnt$pars[((I-1)*(n.pred)+1):length(optnt$pars)]
alpha <- cbind(alpha,1-apply(alpha,1,sum))
```

A.1.2 C++ Code

```
#include <math.h>
```

```
#include<iostream>
#include <TMB.hpp>
// library needed for the multvariate normal distribution
using namespace density;

template<class Type>
vector<Type> Multiply(vector<Type> mat1, matrix<Type> mat2) {

  int i = mat1.size();
  int j = mat2.cols();

  vector<Type> mat3(j);
  for (int c = 0; c < j; c++) {
    mat3(c) = 0.0;
    for (int b = 0; b < i; b++) {
      mat3(c) += mat1(b)*mat2(b,c);
    }
  }
  return(mat3);
}

template<class Type>
matrix<Type> invilrc(matrix<Type> data, matrix<Type> V)
{
  matrix<Type> mult(data.rows(), data.cols()+1);
  matrix<Type> trans(data.rows(),data.cols()+1);
  Type rowtot;

  // back transforming each individual row
  for(int m=0; m<data.rows(); m++){
    rowtot = 0.0;
    for(int l=0; l<(data.cols()+1); l++) mult(m,l) = 0;
```

```
    for(int j=0; j<(data.cols()+1); j++){
      for(int k=0; k<data.cols(); k++)
      {
        mult(m,j) += data(m,k)*V(j,k);
      }
      trans(m,j) = exp(mult(m,j));
      rowtot += trans(m,j);
    }
    trans.row(m) = trans.row(m)/rowtot;
  }
  return(trans);
}

template<class Type>
vector<Type> modzeros(vector<Type> data, double delta)
{
  Type nozero;
  vector<Type> dnozero(data.size());

  //  for(int j=0; j<data.rows(); j++){
  nozero = 0.0;
  for(int k=0; k<data.size(); k++){
    if(data(k)==0.0) nozero += 1.0;
  }

  for(int i=0; i<data.size(); i++){
    if(data(i)==0.0) dnozero(i) = delta;
    else dnozero(i) = (1 - nozero*delta)*data(i);
  }
  return(dnozero);
}
```

```
template<class Type>
matrix<Type> modzerosmat(matrix<Type> data, double delta)
{
  Type nozero;
  matrix<Type> dnozero(data.rows(), data.cols());

  for(int j=0; j<data.rows(); j++){
    nozero = 0.0;
    for(int k=0; k<data.cols(); k++){
      if(data(j,k)==0.0) nozero += 1.0;
    }

    for(int i=0; i<data.cols(); i++){
      if(data(j,i)==0.0) dnozero(j,i) = delta;
      else dnozero(j,i) = (1 - nozero*delta)*data(j,i);
    }
  }
  return(dnozero);
}

template<class Type>
matrix<Type> ilrm(matrix<Type> data, matrix<Type> V)
{
  matrix<Type> trans(data.rows(), data.cols());
  matrix<Type> mult(data.rows(),data.cols()-1);

  Type gmean;
  Type product;

  // going over each row to individually transform rows
  for(int m=0; m<data.rows(); m++){
```

```
    product = 1.0;
    for(int i = 0; i < data.cols(); i++) product *= data(m,i);
    gmean = exp(log(product)/data.cols());

    for(int l=0; l<trans.cols(); l++) trans(m,l) = log(data(m,l)/gmean);

    for(int j=0; j<(trans.cols()-1); j++){
      mult(m,j) = 0.0;
      for(int k=0; k<trans.cols(); k++)
      {
        mult(m,j) += trans(m,k)*V(k,j);
      }
    }
  }
  return(mult);
}

template<class Type>
vector<Type> ilrc(vector<Type> data, matrix<Type> V)
{
  vector<Type> trans(data.size());
  vector<Type> mult(data.size()-1);

  // going over each row to individually transform rows
  // for(int m=0; m<data.rows(); m++){
  Type product = 1.0;
  for(int i = 0; i < data.size(); i++) product *= data(i);
  Type gmean = exp(log(product)/data.size());

  for(int l=0; l<trans.size(); l++) trans(l) = log(data(l)/gmean);

  for(int j=0; j<(trans.size()-1); j++){
```

```
    mult(j) = 0.0;
    for(int k=0; k<trans.size(); k++)
    {
      mult(j) += trans(k)*V(k,j);
    }
  }
  // }
  return(mult);
}

template<class Type>
Type objective_function<Type>::operator() () {

  // Data:
  DATA_MATRIX(y);
  // each row is observed ilr transformed FAs of predators
 // DATA_MATRIX(x);
 // observed ilr transformed prey FAs, each row represents and
 // individual prey FA
  DATA_VECTOR(n);
  // number of observed prey FAs for each specific prey
  // species in x (in same order as in x)
  DATA_MATRIX(varz);
  // estimated pooled variance-covariance matrix of the
  // untransformed prey xi
  DATA_MATRIX(mu);
  // each row is the mean vector for species i (Note this
  // is the mean of the ilr transformed prey FAs)
  DATA_MATRIX(V);
  // V matrix that is D*(D-1) to go from clr transformation
  //to ilr transformation
  // DATA_MATRIX(G);
```

```
 // G matrix that is required to go from T to clr variance
 DATA_VECTOR(sind);
 // start values for the mean quantile of the diagonal
 // of the covariance

 // Parameters:
 // PARAMETER(lambda);
 // Lagrange multiplier parameter
 PARAMETER_MATRIX(alpha);
 // each row is diet proportion vector for predator i, I-1 proportions
 PARAMETER_ARRAY(z);
 // unobserved ilr transformed prey effect, for each
 // individual predator
 PARAMETER_VECTOR(sepsilon);
 // The diagonal entries of the variance-covariance
 // matrix of the ilr transformed epsilon
 // Procedures:
// ADREPORT(alpha);
// ADREPORT(sepsilon);
 //ADREPORT(sepsilon);

 int D = y.cols(); // number of fatty acids - 1 (since ilr transformed)
 int I = n.size(); // number of species
 int npred = y.rows(); // number of predators

 Type nll = 0.0; // initialize negative loglik

 // Latent process:
 vector<Type> tmp(D);
 // initialize point at which to evaluate neg log-density
 // vector<Type> yomean(D+1); // initialize the mean of
 // the predator yo
```

```
matrix<Type> ymean;
// initialize the transformed mean of the predator y

// Rcout << ymean;
array<Type> zarray;
matrix<Type> zo;
matrix<Type> zt;
matrix<Type> eta;

// getting the variance-covariance of z from T
//matrix<Type> varz = V.transpose()*G;
//varz = varz*T;
//varz = varz*G;
//varz = varz*V;
//varz = -0.5*varz;
//REPORT(varz);

MVNORM_t<Type> nll_dist(varz);
// declare multivariate normal with cov mat for prey

// creating a diagonal matrix
matrix<Type> Sigmay(D,D);
Sigmay.fill(0);
for(int k = 0; k < D; k++){
  for(int l=0; l < 4; l++){
    if(sind(k)==(l+1)) Sigmay(k,k) = sepsilon(l);
  }
}

// matrix<Type> Sigmay = sepsilon*Id;
REPORT(Sigmay);
MVNORM_t<Type> nll_y(Sigmay);
```

```
// declare multivariate normal with cov mat for y
//
//

matrix<Type> alphafull(npred,I);
for(int r=0; r<npred; r++){
  vector<Type> vr = alpha.row(r);
  for(int c=0; c<I; c++){
    if(c<I-1) alphafull(r,c) = alpha(r,c);
    else alphafull(r,c) = 1 - sum(vr);
  }
}

for(int w=0; w<npred; w++){
  zarray = z.col(w);
  zt = zarray.matrix();
  //int size = z.cols();
   //REPORT(size);
  zo = invilrc(zt,V);
  //REPORT(zo);
  //
  //eta = Multiply(alpha.row(w),zo);
  eta = alphafull.row(w)*zo;
  REPORT(eta);

  eta = modzerosmat(eta, 0.00005);
  ymean = ilrm(eta,V);
  REPORT(ymean);

  //   // Observation models:
  tmp = y.row(w) - ymean; // centers it
  nll += nll_y(tmp);
```

```
    //  adding in the likelihood of z, the random effect
    for(int i = 0; i < I; i++){ //
      tmp =  zt.row(i) - mu.row(i); // centers it
      nll += nll_dist(tmp);
    }
  }




  //
  //REPORT(nll);
  return nll;
}
```

A.2 Bootstraps

A.2.1 R Code

```
gen.est <- function(npred, preym, preymt, poolvar, diet, D,
 preybaseq, FAs, V,G, sepstrue=rep(0.001,D-1),
  sindtrue = NA){

# Simulating from the true diet
sim.mat <- matrix(NA, nrow=npred, ncol=ncol(preym))
if(is.matrix(diet)){
  for(i in 1:n.pred){
    sim.mat[i,] <- pseudo.seal.norm(preymt, poolvar, diet[i,])
  }
}else
for(i in 1:n.pred){
```

```
  sim.mat[i,] <- pseudo.seal.norm(preymt, poolvar, diet)
}

if(is.na(sindtrue)){
  e <- rmvnorm(nrow(sim.mat), rep(0,D-1), diag(sepstrue))

}else{
  diagseps <- numeric(D-1)
  for(j in 1:4){
    diagseps[sindtrue==j] <- sepstrue[j]
  }
  e <- rmvnorm(nrow(sim.mat), rep(0,D-1), diag(diagseps))
}

e0 <- acomp(ilrInv(e))

sim.use <- perturbe(sim.mat,e0)
colnames(sim.use) <- colnames(preybaseq)[-1]

test <- p.QFASA(predator.mat=sim.use,prey.mat=preym,
 cal.mat=matrix(rep(1,npred*I),ncol=npred), dist.meas=2,
   start.val=rep(1/I,I), ext.fa=FAs )

start.val <- test$`Diet Estimates`[,-I]

ers <- matrix(NA, nrow = 50*npred, ncol = ncol(sim.use))

for(j in 1:50){
  yest <- matrix(NA, nrow = npred, ncol = ncol(sim.use))
  for(i in 1:nrow(yest)){
    yest[i,] <- pseudo.seal.norm(preymt, poolvar,
     test$`Diet Estimates`[i,])
```

```
  }

  yest <- acomp(yest) #without error, untransformed
  lb <- (j-1)*npred + 1
  ub <- j*npred
  ers[lb:ub,] = -yest + sim.use
}
print(ers)
ers <- acomp(ers)
sep.start <- diag(-1/2*t(V)%*%G%*%variation(ers)%*%G%*%V)

quan <- quantile(sep.start)
quan.start <- numeric(4)
groupind <- numeric(length(sep.start))
for(j in 1:4){
  quan.start[j] <- mean(sep.start[sep.start>=quan[j] &
    sep.start<=quan[j+1]])
  groupind[sep.start>=quan[j] & sep.start<=quan[j+1]] <- j
}

pred <- mod.zeros.FA.sig.mat(sim.use,0.00005)
#print(pred)
ilr.data <- ilr(pred)

parameters <- list(alpha = start.val,
                   z  = array(rep(preymt), c(nrow(preymt),
                     ncol(preymt), npred)),
                   sepsilon = quan.start
)

#Data to send to tmb
data <- list(y = ilr.data,
```

```
              #             x = preybase.t,
              n=n,
              varz=poolvar,
              mu = preymt,
              V=ilrBase(D=(ncol(ilr.data)+1)),
              sind=groupind
)

objnt <- MakeADFun(data,parameters,random="z")
npars <- length(objnt$par)

# constraining alphas to be between 0 and 1
lb <- rep(0,npars)
ub <- c(rep(1,(npars-length(quan.start))), rep(Inf,length(quan.start)))



# Full parameter set
al.sum <- function(pars){
  npars <- length(pars)
  alpha <- matrix(pars[1:(npars-length(quan.start))], ncol=I-1)
  return(apply(alpha,1,sum))
}

# Full parameter set
optnt <- solnp(pars=objnt$par, fun=objnt$fn, ineqfun=al.sum,
ineqLB = rep(0,npred), ineqUB = rep(1,npred), LB=lb,
UB=ub, control=list(delta=0.0001, tol=0.00001))

L <- optnt$values[length(optnt$values)]

alpha <- matrix(optnt$pars[1:((I-1)*(n.pred))],nrow=n.pred)
seps <- optnt$pars[((I-1)*(n.pred)+1):length(optnt$pars)]
```

```
alpha <- cbind(alpha,1-apply(alpha,1,sum))
conv <- optnt$convergence

return(list(L=L,alpha=alpha,seps=seps, sind=groupind, conv=conv))
}

##### Use these estimates as start values
#preymt =prey.mt
#poolvar = Spool
#preym = prey.m
#sind = groupind

# MLEvals is a list with L=negative log likelihood value,
#alpha=diet estimates
# obtained from ML method, seps = epsilon diagonal obtained from
# ML method, sind = indicator of the diagonal of epsilon
# for where each estimate goes, conv = 0 for convergence

boots.run <- function(npred, preym, preymt, poolvar,
diet, D, preybaseq, FAs, V, G, sepstrue, r, sindtrue=NA){

   for(i in 1:r){
     run.r <- gen.est(npred, preym, preymt,
       poolvar,diet,D,preybaseq,FAs, V,G,sepstrue, sindtrue)
     als <- run.r$alpha
     seps <- run.r$seps
     Ls <- run.r$L

     if(i==1){
       Lr <- Ls
       alr <- als
       epsr <- seps
```

```
      conv <- run.r$conv
    }else{
      Lr <- c(Lr,Ls)
      alr <- rbind(alr, als)
      epsr <- rbind(epsr, seps)
      conv <- c(conv,run.r$conv)
    }

  }
 return(list(L=Lr, alpha = alr, seps=epsr, conv=conv))
}


library(doMC)
registerDoMC(cores=5)
library(foreach)
bootset <- foreach(r = rep(20,5)) %dopar%
  boots.run(n.pred, prey.m, prey.mt, Spool, MLEvals$alpha, D,
    preybase.q, FAset, V, G, sepstrue=MLEvals$seps,r,
      sindtrue=MLEvals$sind  )
```

A.3 Covariates

A.3.1 R Code

```
source("est.functions.R")

#FAset to use
FAset <- read.csv("DietFAPhocidGreys.csv")
FAset <- as.vector(unlist(FAset))

## Preybase to use
```

```
preybase <- read.csv("SableHgModelPreySetFallWinter.csv")
# No 16:4w3 for greys
preybase <- preybase[,!(colnames(preybase)=="c16.4w3")]

species <- unique(preybase$SableHgModelGroupNameFallWinter)
species <- sort(species)
I <- length(species)

# Extracting and ordering prey species
for(j in 1:length(species)){
   if(j==1) {preybase.spec
    <- preybase[preybase$SableHgModelGroupNameFallWinter==species[j],] }
   else{
      preybase.spec <- rbind(preybase.spec,
       preybase[preybase$SableHgModelGroupNameFallWinter==species[j],])
   }
}
preybase <- preybase.spec

# Winsorizing (optional)
preybase.w <- matrix(NA,nrow=nrow(preybase), ncol=(ncol(preybase)-5))
rownames(preybase.w) <- preybase$SableHgModelGroupNameFallWinter

for(i in 1:I){
  preyi
   <- preybase[preybase$SableHgModelGroupNameFallWinter==species[i],][,-(1:5)]
  for(j in 1:(ncol(preybase)-5)){
    iqr <- IQR(preyi[,j])
    vals <- quantile(preyi[,j], c(0.25, 0.5, 0.75))
    low <- vals[1] - 1.5*iqr
    high <- vals[3] + 1.5*iqr
    preybase.w[rownames(preybase.w)==species[i],j]
```

```
      <- Winsorize(preyi[,j], minval = low, maxval=high)
  }
}

colnames(preybase.w) <- colnames(preybase[,-(1:5)])


preybase.q <- preybase.w


# Extract FAs from prey data base
preybase.q <- preybase.q[,FAset]
preybase.q <- preybase.q/apply(preybase.q,1,sum)
sort.preytype <- order(rownames(preybase.q))
preybase.q <- preybase.q[sort.preytype,]


# Transforming full prey base
preybase.t <- mod.zeros.FA.sig.mat(preybase.q,0.00005)
 #analytic precision is two decimals after the percentage.
rownames(preybase.t) <- rownames(preybase.q)
preybase.t <-ilr(preybase.t)


## Seals
seals.full <- read.csv("SealFASignaturesMay2020.csv", header=TRUE)
seals <- seals.full[,FAset]/rowSums(seals.full[,FAset])

npred <- nrow(seals)

## Calibration Coefficients
cals.orig <- read.csv("HgCCs.csv")
```

```
cals <- cals.orig$GreySealCC
names(cals) <- cals.orig$FA
cals <- cals[FAset]


# getting the number sampled from each species
I <- length(species)
n <- tapply(rownames(preybase.q), rownames(preybase.q), length)

## Estimating the variance covariance matrices of our prey data
prey.var <- vector("list", I)
prey.mt <- matrix(0, nrow=I, ncol=length(FAset)-1)
prey.m <- matrix(0, nrow=I, ncol=length(FAset))
for(i in 1:I){
  preyi <- preybase.q[rownames(preybase.q)==species[i],]
  preyit <- preybase.t[rownames(preybase.t)==species[i],]
  prey.var[[i]] <- variation.acomp(preyi)
  prey.mt[i,] <- apply(preyit,2,mean)
  prey.m[i,] <- apply(preyi,2,mean)
}
D <- length(FAset)
V <- ilrBase(D=length(FAset))
G <- V%*%t(V)

## Calculating the pooled variance estimate
Spool <- matrix(0, D-1,D-1)
for(j in 1:I){
  Sigma <- -0.5*t(V)%*%G%*%prey.var[[j]]%*%G%*%V
  Spool <- Spool + (n[[j]]-1)*Sigma
}
Spool <- Spool/(sum(n) - length(n))
```

```
### create and load function from .cpp template
compile("ErrorCovariates.cpp", flags="-Wno-unused-variable")
# create .o and .so
#dyn.unload(dynlib("ErrorCovariates"))
dyn.load(dynlib("ErrorCovariates")) # load .so


# Sex and Year Group
 X <- cbind(rep(1,nrow(seals.full)),dummy(seals.full$Sex),
   dummy(seals.full$YearGroup))

# Estimation function
 cov.est <- function(preds, preym, calmat, FAs, preymt, poolvar,
   D, V,G, X){

   test <- p.QFASA(predator.mat=unclass(preds),prey.mat=preym,
     cal.mat=calmat, dist.meas=2, start.val=rep(1/I,I),
      ext.fa=FAs )

   start.val <- test$`Diet Estimates`

   # Seals after calibration coefficients
   Sealsc <- matrix(NA,nrow=nrow(preds), ncol=ncol(preds))
   for(i in 1:length(test$`Additional Measures`)){
     Sealsc[i,] <- test$`Additional Measures`[[i]]$ModFAS
   }

  # Start values for betas
   start.beta <- atob(X,start.val)

   npred <- nrow(preds)
```

```
ers <- matrix(NA, nrow = 50*npred, ncol = ncol(preds))

for(j in 1:50){
  yest <- matrix(NA, nrow = npred, ncol = ncol(preds))
  for(i in 1:nrow(yest)){
    yest[i,] <- pseudo.seal.norm(preymt, poolvar,
     test$`Diet Estimates`[i,])
  }

  yest <- acomp(yest) #without error, untransformed
  lb <- (j-1)*npred + 1
  ub <- j*npred
  ers[lb:ub,] = -yest + preds
}
print(ers)
ers <- acomp(ers)
sep.start <- diag(-1/2*t(V)%*%G%*%variation(ers)%*%G%*%V)


quan <- quantile(sep.start)
quan.start <- numeric(4)
groupind <- numeric(length(sep.start))
for(j in 1:4){
  quan.start[j] <- mean(sep.start[sep.start>=quan[j] &
    sep.start<=quan[j+1]])
  groupind[sep.start>=quan[j] & sep.start<=quan[j+1]] <- j
}


pred <- mod.zeros.FA.sig.mat(Sealsc,0.00005)
#print(pred)
```

```
   ilr.data <- ilr(pred)

# Parameters to be passed to TMB
   parameters <- list(beta = start.beta
                      z  = array(rep(prey.mt), c(nrow(prey.mt),
                        ncol(prey.mt), npred)),
                      sepsilon = quan.start
   )

   #Data to send to tmb
   data <- list(y = ilr.data,
                #            x = preybase.t,
                n=n,
                varz=Spool,
                mu = prey.mt,
                V=ilrBase(D=(ncol(ilr.data)+1)),
                Xcov= X,
                sind=groupind
   )

   objnt <- MakeADFun(data,parameters,random="z")
   npars <- length(objnt$par)


   lb <- c(rep(-Inf,npars-4), rep(0,4))
   ub <- rep(Inf,npars)

   optnt <- nlminb(start=objnt$par, objective=objnt$fn, lower=lb,
     upper=ub) #control=list(abs.tol=0.001, iter.max = 1,
       eval.max=1000)

   L <- optnt$objective
```

```
   betaopt <- matrix(optnt$par[1:(npars-4)], nrow=((npars-4)/(I-1)))
   alphaopt <- btoa(X,betaopt)
   seps <- optnt$par[(length(optnt$par)-3):length(optnt$par)]
   conv <- optnt$convergence
   starts <- start.val

   return(list(L=L,beta=betaopt,seps=seps, conv=conv, Q=starts,
     groupind=groupind))
 }


MLEvals <- cov.est(as.matrix(seals), prey.m, cals, FAset,
 prey.mt, Spool, D, V, G, X)
```

A.3.2 C++ Code

```
#include <math.h>
// Including covariates, first simple just male and female
#include<iostream>
#include <TMB.hpp>
// library needed for the multvariate normal distribution
using namespace density;


template<class Type>
matrix<Type> Beta2Alpha(matrix<Type> mat1, matrix<Type> mat2) {

  int i = mat1.rows(); // this is number of predators
  int j = mat2.cols(); // this is number of prey species I-1

  matrix<Type> mat3 = mat1 * mat2; // multiplying X and B
```

```
  matrix<Type> al(i,(j+1)); // alpha is nxI
  for(int r=0; r<i; r++){ // over the rows
    //Type alsum = 0.0;
    al(r,0) = exp(0);// first column is 1s
    //alsum = al(r,0);
    for(int c=1; c<(j+1); c++){ // over I, the number of prey species
      al(r,c) = exp(mat3(r,(c-1)));
     // alsum += al(r,c);
    }
    al.row(r) = al.row(r)/al.rowwise().sum()(r);
    // divide by sums so alpha is compositional
  }

  return(al);
}



template<class Type>
matrix<Type> invilrc(matrix<Type> data, matrix<Type> V)
{
  matrix<Type> mult(data.rows(), data.cols()+1);
  matrix<Type> trans(data.rows(),data.cols()+1);
  Type rowtot;

  // back transforming each individual row
  for(int m=0; m<data.rows(); m++){
    rowtot = 0.0;
    for(int l=0; l<(data.cols()+1); l++) mult(m,l) = 0;

    for(int j=0; j<(data.cols()+1); j++){
      for(int k=0; k<data.cols(); k++)
      {
```

```
        mult(m,j) += data(m,k)*V(j,k);
      }
      trans(m,j) = exp(mult(m,j));
      rowtot += trans(m,j);
    }
    trans.row(m) = trans.row(m)/rowtot;
  }
  return(trans);
}

template<class Type>
vector<Type> modzeros(vector<Type> data, double delta)
{
  Type nozero;
  vector<Type> dnozero(data.size());

  //  for(int j=0; j<data.rows(); j++){
  nozero = 0.0;
  for(int k=0; k<data.size(); k++){
    if(data(k)==0.0) nozero += 1.0;
  }

  for(int i=0; i<data.size(); i++){
    if(data(i)==0.0) dnozero(i) = delta;
    else dnozero(i) = (1 - nozero*delta)*data(i);
  }
  return(dnozero);
}

template<class Type>
matrix<Type> modzerosmat(matrix<Type> data, double delta)
{
```

```
  Type nozero;
  matrix<Type> dnozero(data.rows(), data.cols());

  for(int j=0; j<data.rows(); j++){
    nozero = 0.0;
    for(int k=0; k<data.cols(); k++){
      if(data(j,k)==0.0) nozero += 1.0;
    }

    for(int i=0; i<data.cols(); i++){
      if(data(j,i)==0.0) dnozero(j,i) = delta;
      else dnozero(j,i) = (1 - nozero*delta)*data(j,i);
    }
  }
  return(dnozero);
}

template<class Type>
matrix<Type> ilrm(matrix<Type> data, matrix<Type> V)
{
  matrix<Type> trans(data.rows(), data.cols());
  matrix<Type> mult(data.rows(),data.cols()-1);

  Type gmean;
  Type product;

  // going over each row to individually transform rows
  for(int m=0; m<data.rows(); m++){
    product = 1.0;
    for(int i = 0; i < data.cols(); i++) product *= data(m,i);
    gmean = exp(log(product)/data.cols());
```

```
    for(int l=0; l<trans.cols(); l++) trans(m,l) = log(data(m,l)/gmean);

    for(int j=0; j<(trans.cols()-1); j++){
      mult(m,j) = 0.0;
      for(int k=0; k<trans.cols(); k++)
      {
        mult(m,j) += trans(m,k)*V(k,j);
      }
    }
  }
  return(mult);
}

template<class Type>
vector<Type> ilrc(vector<Type> data, matrix<Type> V)
{
  vector<Type> trans(data.size());
  vector<Type> mult(data.size()-1);

  // going over each row to individually transform rows
  // for(int m=0; m<data.rows(); m++){
  Type product = 1.0;
  for(int i = 0; i < data.size(); i++) product *= data(i);
  Type gmean = exp(log(product)/data.size());

  for(int l=0; l<trans.size(); l++) trans(l) = log(data(l)/gmean);

  for(int j=0; j<(trans.size()-1); j++){
    mult(j) = 0.0;
    for(int k=0; k<trans.size(); k++)
    {
      mult(j) += trans(k)*V(k,j);
```

```
    }
  }
  // }
  return(mult);
}

template<class Type>
Type objective_function<Type>::operator() () {

  // Data:
  DATA_MATRIX(y);
  // each row is observed ilr transformed FAs of predators
  DATA_VECTOR(n);
  // number of observed prey FAs for each specific prey
  // species in x (in same order as in x)
  DATA_MATRIX(varz);
  // estimated pooled variance-covariance matrix of
  // the untransformed prey xi
  DATA_MATRIX(mu);
  // each row is the mean vector for species i
  // (Note this is the mean of the ilr transformed prey FAs)
  DATA_MATRIX(V);
  // V matrix that is D*(D-1) to go from clr transformation
  // to ilr transformation
  // DATA_MATRIX(G);
   // G matrix that is required to go from T to clr variance
  DATA_MATRIX(Xcov);
  // Matrix npred*(p+1), first column 1s, column 2
  // covariate 1, etc
  DATA_VECTOR(sind);
  // start values for the mean quantile of the diagonal
  // of the covariance
```

```
  // Parameters:
  // PARAMETER(lambda);
  // Lagrange multiplier parameter
  PARAMETER_MATRIX(beta);
  // Coefficients matrix, (p+1)*(I-1)
  PARAMETER_ARRAY(z);
  // unobserved ilr transformed prey effect, for each individual predator
  PARAMETER_VECTOR(sepsilon);
  // The diagonal entries of the variance-covariance
  // matrix of the ilr transformed epsilon
  // Procedures:
  // ADREPORT(sepsilon);
  //ADREPORT(sepsilon);

  matrix<Type> alpha = Beta2Alpha(Xcov,beta);
  // REPORT(alpha);
   // REPORT(beta);
  // REPORT(Xcov);
  //
  // ADREPORT(alpha);

  int D = y.cols(); // number of fatty acids - 1 (since ilr transformed)
  int I = n.size(); // number of species
  int npred = y.rows(); // number of predators

   Type nll = 0.0; // initialize negative loglik

// Latent process:
  vector<Type> tmp(D);
  // initialize point at which to evaluate neg log-density
```

```
 // Rcout << ymean;
 array<Type> zarray;
 matrix<Type> zo;
 matrix<Type> zt;
// matrix<Type> eta;



 MVNORM_t<Type> nll_dist(varz);
 // declare multivariate normal with cov mat for prey

 // creating a diagonal matrix
 matrix<Type> Sigmay(D,D);
 Sigmay.fill(0);
 for(int k = 0; k < D; k++){
   for(int l=0; l < 4; l++){
     if(sind(k)==(l+1)) Sigmay(k,k) = sepsilon(l);
   }
 }


 MVNORM_t<Type> nll_y(Sigmay);
 // declare multivariate normal with cov mat for y
 //


 for(int w=0; w<npred; w++){
   zarray = z.col(w);
   zt = zarray.matrix();

   zo = invilrc(zt,V);
```

```
    matrix<Type> eta = alpha.row(w)*zo;
    //REPORT(eta);

    eta = modzerosmat(eta, 0.00005);
    matrix<Type> ymean = ilrm(eta,V);
   // REPORT(ymean);

    // Observation models:
    tmp = y.row(w) - ymean; // centers it
    nll += nll_y(tmp);

    //  adding in the likelihood of z, the random effect
    for(int i = 0; i < I; i++){ //
      tmp =  zt.row(i) - mu.row(i); // centers it
      nll += nll_dist(tmp);
    }
  }


  return nll;
}
```

A.4 Inference

A.4.1 R Code

```
source("est.functions.R")

#FAset to use
```

```
FAset <- read.csv("DietFAPhocidGreys.csv")
FAset <- as.vector(unlist(FAset))

## Preybase to use
preybase <- read.csv("SableHgModelPreySetFallWinter.csv")
# No 16:4w3 for greys
preybase <- preybase[,!(colnames(preybase)=="c16.4w3")]

species <- unique(preybase$SableHgModelGroupNameFallWinter)
species <- sort(species)
I <- length(species)

for(j in 1:length(species)){
  if(j==1) {preybase.spec <-
   preybase[preybase$SableHgModelGroupNameFallWinter==species[j],]}
  else{
    preybase.spec <- rbind(preybase.spec,
     preybase[preybase$SableHgModelGroupNameFallWinter==species[j],])
  }
}

preybase <- preybase.spec

preybase.w <- matrix(NA,nrow=nrow(preybase), ncol=(ncol(preybase)-5))
rownames(preybase.w) <- preybase$SableHgModelGroupNameFallWinter

for(i in 1:I){
  preyi <- preybase[preybase$SableHgModelGroupNameFallWinter
   ==species[i],][,-(1:5)]
  for(j in 1:(ncol(preybase)-5)){
    iqr <- IQR(preyi[,j])
    vals <- quantile(preyi[,j], c(0.25, 0.5, 0.75))
```

```
    low <- vals[1] - 1.5*iqr
    high <- vals[3] + 1.5*iqr
    preybase.w[rownames(preybase.w)==species[i],j]
      <- Winsorize(preyi[,j], minval = low, maxval=high)
  }
}

colnames(preybase.w) <- colnames(preybase[,-(1:5)])

preybase.q <- preybase.w


# Extract FAs from prey data base
preybase.q <- preybase.q[,FAset]
preybase.q <- preybase.q/apply(preybase.q,1,sum)
sort.preytype <- order(rownames(preybase.q))
preybase.q <- preybase.q[sort.preytype,]


# Transforming full prey base
preybase.t <- mod.zeros.FA.sig.mat(preybase.q,0.00005)
#analytic precision is two decimals after the percentage.
rownames(preybase.t) <- rownames(preybase.q)
preybase.t <-ilr(preybase.t)


## Seals

seals.full <- read.csv("SealFASignaturesMay2020.csv", header=TRUE)

seals <- seals.full[,FAset]/rowSums(seals.full[,FAset])
seals <- as.matrix(seals)
```

```
npred <- nrow(seals)

## Calibration Coefficients
cals.orig <- read.csv("HgCCs.csv")
cals <- cals.orig$GreySealCC
names(cals) <- cals.orig$FA
cals <- cals[FAset]



# getting the number sampled from each species
I <- length(species)
n <- tapply(rownames(preybase.q), rownames(preybase.q), length)

## Estimating the variance covariance matrices of our prey data
prey.var <- vector("list", I)
prey.mt <- matrix(0, nrow=I, ncol=length(FAset)-1)
prey.m <- matrix(0, nrow=I, ncol=length(FAset))
for(i in 1:I){
  preyi <- preybase.q[rownames(preybase.q)==species[i],]
  preyit <- preybase.t[rownames(preybase.t)==species[i],]
  prey.var[[i]] <- variation.acomp(preyi)
  prey.mt[i,] <- apply(preyit,2,mean)
  prey.m[i,] <- apply(preyi,2,mean)
}
D <- length(FAset)
V <- ilrBase(D=length(FAset))
G <- V%*%t(V)

## Calculating the pooled variance estimate
Spool <- matrix(0, D-1,D-1)
for(j in 1:I){
```

```
  Sigma <- -0.5*t(V)%*%G%*%prey.var[[j]]%*%G%*%V
  Spool <- Spool + (n[[j]]-1)*Sigma
}
Spool <- Spool/(sum(n) - length(n))

### create and load function from .cpp template
compile("ErrorCovariates.cpp", flags="-Wno-unused-variable")
 # create .o and .so
#dyn.unload(dynlib("ErrorCovariates"))
dyn.load(dynlib("ErrorCovariates")) # load .so

#Covariate matrix

# Sex and Year Group
X <- cbind(rep(1,nrow(seals.full)),dummy(seals.full$Sex))

# I is the number of prey species
I <- nrow(prey.mt)

########################################

np <- nrow(seals)

cov.est <- function(preds, preym, calmat, FAs, preymt, poolvar,
 D, V,G, X){

  test <- p.QFASA(predator.mat=unclass(preds),prey.mat=preym,
    cal.mat=calmat, dist.meas=2, start.val=rep(1/I,I),
     ext.fa=FAs )

  start.val <- test$‘Diet Estimates‘
```

```
npred <- nrow(preds)

# Seals after calibration coefficients
Sealsc <- matrix(NA,nrow=nrow(preds), ncol=ncol(preds))
for(i in 1:length(test$`Additional Measures`)){
  Sealsc[i,] <- test$`Additional Measures`[[i]]$ModFAS
}



# Starting values for betas
#reg <- zadr(start.val,X[,-1],xnew=X[,-1])
#start.beta <- reg$be
start.beta <- atob(X,start.val)

ers <- matrix(NA, nrow = 50*npred, ncol = ncol(preds))

for(j in 1:50){
  yest <- matrix(NA, nrow = npred, ncol = ncol(preds))
  for(i in 1:nrow(yest)){
    yest[i,] <- pseudo.seal.norm(preymt, poolvar, start.val[i,])
  }

  yest <- acomp(yest) #without error, untransformed
  lb <- (j-1)*npred + 1
  ub <- j*npred
  ers[lb:ub,] = -yest + preds
}
print(ers)
ers <- acomp(ers)
sep.start <- diag(-1/2*t(V)%*%G%*%variation(ers)%*%G%*%V)
```

```
quan <- quantile(sep.start)
quan.start <- numeric(4)
groupind <- numeric(length(sep.start))
for(j in 1:4){
  quan.start[j] <- mean(sep.start[sep.start>=quan[j] &
    sep.start<=quan[j+1]])
  groupind[sep.start>=quan[j] & sep.start<=quan[j+1]] <- j
}



pred <- mod.zeros.FA.sig.mat(Sealsc,0.00005)
#print(pred)
ilr.data <- ilr(pred)

parameters <- list(beta = start.beta,
                   z  = array(rep(prey.mt), c(nrow(prey.mt),
                     ncol(prey.mt), npred)),
                   sepsilon = quan.start
)

#Data to send to tmb
data <- list(y = ilr.data,
             #            x = preybase.t,
             n=n,
             varz=Spool,
             mu = prey.mt,
             V=ilrBase(D=(ncol(ilr.data)+1)),
             Xcov= X,
             sind=groupind
)

objnt <- MakeADFun(data,parameters,random="z")
```

```
  npars <- length(objnt$par)
  I <- nrow(preym)


  lb <- c(rep(-Inf,npars-4), rep(0,4))
  ub <- rep(Inf,npars)

  optnt <- nlminb(start=objnt$par, objective=objnt$fn, lower=lb,
    upper=ub)

  L <- optnt$objective
  betaopt <- matrix(optnt$par[1:(npars-4)], nrow=((npars-4)/(I-1)))
  alphaopt <- btoa(X,betaopt)
  seps <- optnt$par[(npars-3):npars]
  conv <- optnt$convergence
  starts <- start.val

  return(list(L=L,beta=betaopt,seps=seps, conv=conv, Q=starts,
    groupind=groupind))
}

MLEval <- cov.est(seals, prey.m, cals, FAset, prey.mt, Spool, D, V,G, X)
n.pred <- nrow(seals)

 groupind <- MLEval$groupind
 diageps <- numeric(length(groupind))
 for(j in 1:4){
   diageps[groupind==j] <- MLEval$seps[j]
 }

bootsbeta <- function(betaopt, X, npred, preym, cal.mat, preymt,
 poolvar, D, preybaseq, FAs, V, G, diageps, r){
```

```
diet <- btoa(X,betaopt)
for(i in 1:r){

  #pseudo-predators with true values of MLE diets
  sim.mat <- matrix(NA, nrow=npred, ncol=ncol(preym))
  for(j in 1:npred){
    sim.mat[j,] <- pseudo.seal.norm(preymt, poolvar, diet[j,])
  }

  e <- rmvnorm(nrow(sim.mat), rep(0,D-1), diag(diageps))

  #
  e0 <- acomp(ilrInv(e))

  sim.use <- perturbe(sim.mat,e0)
  colnames(sim.use) <- colnames(preybaseq)

  I <- nrow(preym)
  seals <- as.matrix(sim.use)

  simset <- cov.est(seals, prey.m, cals, FAset, prey.mt, Spool, D, V,
            G, X)

  als <- as.vector(betaopt[-1,])
  seps <- seps
  Ls <- L

  if(i==1){
    Lr <- Ls
    alr <- als
    epsr <- seps
```

```
      conv <- convr
    }else{
      Lr <- c(Lr,Ls)
      alr <- rbind(alr, als)
      epsr <- rbind(epsr, seps)
      conv <- c(conv,convr)
    }

    gc()
  }
  return(list(L=Lr, beta = alr, seps=epsr, conv=conv))
}

library(doMC)
registerDoMC(cores=10)
library(foreach)
simset <- foreach(n.sim = rep(5,10)) %dopar%
  bootsbeta(MLEvalbeta, X, n.pred, prey.m, cals, prey.mt, Spool,
    D, preybase.q,
            FAset, V, G, diageps, n.sim )

beta <- boots$beta

Tbar <- apply(beta,2,mean)
st <- cov(beta)

mdists <- mahalanobis(beta, center=Tbar, cov=st)
Dub <- quantile(mdists, 0.95)

D02 <- (t(Tbar))%*%solve(st)%*%t(t(Tbar))

## Fail to reject if D02 < Dub
```

A.5 est.functions

```
## Functions script for comparison study
## Created: Jan 25, 2017



## The next two functions are used for all methods.
# They replace the zeros in the
## FA signatures
mod.zeros.FA.sig.mat <- function(y.mat,delta) {

  # JANUARY 24TH, 2014
  # MODIFIES THE ZEROS IN A SAMPLE OF FA SIGNATURES USING THE
  # MULTIPLICATIVE REPLACEMENT
  # STRATEGY (AND THE SAME delta FOR EVERY ZERO)

  y.mat <- t(apply(y.mat,1,mod.zeros.FA.sig,delta=delta))

  return(y.mat)
}

mod.zeros.FA.sig <- function(y,delta) {

  # JANUARY 24TH, 2014
  # MODIFIES THE ZEROS IN A SINGLE FA SIGNATURES USING THE
  # MULTIPLICATIVE REPLACEMENT
  # STRATEGY
```

```
  no.zero <- sum(y==0)
  y[y>0] <- (1 - no.zero*delta)*y[y>0]
  y[y == 0] <- delta

  return(y)
}

## The following function is the code for pseudo-seal generation.
#It is based on the
## assumption that the ilr transformations are normally
#distributed and the FA of the
## Predator is the linear combination of diet estimates
#and Prey FAs on the untransformed
## scale. This is the more complicated model.

pseudo.seal.norm <- function(mu.mat, sigma.pool, diet){
  # mu.mat = matrix where each row represents the mean
  #transformed fatty acid signature of each prey type
  # sigma.pool = pooled variance-covariance matrix of the
  # transformed fatty acid signatures of prey types
  # diet = vector of proportions of prey species in diet (true diet)
  J <- length(diet)

  x.mat <- matrix(0, nrow=J, ncol=ncol(mu.mat))
  for (j in 1:J){
    x.mat[j,] <- mvrnorm(1, mu=mu.mat[j,], Sigma=sigma.pool)
  }

  x.mat.o <- ilrInv(x.mat)
  x.mat.o <- as.data.frame(x.mat.o)
  x.mat.o <- as.matrix(x.mat.o)
  Y <- diet %*% x.mat.o
```

```
  return(Y)

}



pseudo.seal.nonparam <- function(prey.mat, species, diet){

  # diet = vector of proportions of prey species in diet (true diet)
  J <- length(diet)

  x.mat <- matrix(0, nrow=J, ncol=ncol(prey.mat))
  for (j in 1:J){
    prey.spec <- prey.mat[rownames(prey.mat) == species[j],]
    x.mat[j,] <- apply(prey.spec[sample(1:nrow(prey.spec),
     nrow(prey.spec), replace=TRUE),],2,mean)
  }

  #x.mat.o <- ilrInv(x.mat)
  #x.mat.o <- as.data.frame(x.mat.o)
  #x.mat.o <- as.matrix(x.mat.o)
  Y <- diet %*% x.mat
  return(Y)

}

## This function computes Aitchison's mean
a.mean <- function(comp.mat){
  gmean <- apply(comp.mat, 2, prod)
  gmean <- exp(log(gmean)/nrow(comp.mat))
  gmean <- gmean/sum(gmean)
  return(gmean)
```

```
}

MEANmeth <- function(prey.mat) {

  # RETURNS THE MULTIVARIATE MEAN FA SIGNATURE FROM EACH PREY TYPE
  # RESULT CAN BE PASSED TO prey.mat IN p.QFASA

  # INPUT:
  # prey.mat --> MATRIX CONTAINING FA SIGNATURES OF THE PREY.  NOTE THAT
  #   FIRST COLUMN INDEXES PREY TYPE.

  prey.means <- apply(prey.mat[, -1], 2, tapply, prey.mat[, 1], mean)
  return(prey.means)
}


AIT.obj <- function(alpha, seal, prey.quantiles) {


  # AUGUST 12TH, 2014
  # SIMILAR TO optcompdiff.obj BUT DOES NOT NORMALIZE ALPHA.

  # USED IN solnp AS THE OBJECTIVE FUNCTION TO BE
  # MINIMIZED

  # INPUT:
  # alpha  --> VECTOR OVER WHICH MINIMIZATION TAKES PLACE
  # seal  --> VECTOR OF  FATTY ACID COMPOSITIONS OF SEAL
  # prey.quantiles  --> MATRIX OF FATTY ACID COMPOSITION OF PREY.
  #     EACH ROW CONTAINS AN INDIVIDUAL PREY
  #    FROM A DIFFERENT SPECIES.
```

```
  no.zero <- sum(seal == 0.)
  seal[seal == 0.] <- 1e-05
  seal[seal > 0.] <- (1. - no.zero * 1e-05) * seal[seal > 0.]

  sealhat <- t(as.matrix(alpha)) %*% prey.quantiles
  no.zero <- sum(sealhat == 0.)
  sealhat[sealhat == 0.] <- 1e-05
  sealhat[sealhat > 0.] <- (1. - no.zero * 1e-05) * sealhat[sealhat > 0.]



  return(AIT.dist(seal, sealhat))
}

AIT.dist <- function(x, bigX) {

  # NOTE THAT THIS FUNCTION IS DIFFERENT THAT compdiff
  #BECAUSE IT TAKES SQUARE ROOT
  # compdiff IS USED IN CALCULATING DIET ESTIMATES.

  # COMPUTES THE DIFFERENCE BETWEEN TO VECTORS OF
  #COMPOSITIONAL DATA
  # AS DESCRIBED IN AITCHISON (1992) "MEASURES OF
  #COMPOSITIONAL DIFFERENCE"

  return(sqrt(sum((log(x/mean.geometric(x)) -
   log(bigX/mean.geometric(bigX)))^2.)))
}

AIT.more <- function(alpha, seal, prey.quantiles) {
```

```
# OCTOBER 28TH, 2014
# USED TO PROVIDE ADDITIONAL INFORMATION ON MODEL COMPONENTS WHEN
# alpha CORRESPONDS TO THE QFASA DIET ESTIMATES (i.e. ESTIMATES
# THAT MINIMIZED THE AIT DISTANCE.)

# THE OBJECTIVE FUNCTION TO BE MINIMIZED

# INPUT:
# alpha                --> VECTOR OVER WHICH MINIMIZATION TAKES PLACE
# seal                  --> VECTOR OF  FATTY ACID COMPOSITIONS OF SEAL
# prey.quantiles  --> MATRIX OF FATTY ACID COMPOSITION OF PREY.
#                                EACH ROW CONTAINS AN INDIVIDUAL PREY
#                                FROM A DIFFERENT SPECIES.


no.zero <- sum(seal == 0.)
seal[seal == 0.] <- 1e-05
seal[seal > 0.] <- (1. - no.zero * 1e-05) * seal[seal > 0.]

sealhat <- t(as.matrix(alpha)) %*% prey.quantiles
no.zero <- sum(sealhat == 0.)
sealhat[sealhat == 0.] <- 1e-05
sealhat[sealhat > 0.] <- (1. - no.zero * 1e-05) * sealhat[sealhat > 0.]

AIT.sq.vec <-
  ( log(seal/mean.geometric(seal)) -
  log(sealhat/mean.geometric(sealhat)) )^2

dist <- (sum(AIT.sq.vec))^(1/2)

out.list <- list(sealhat,AIT.sq.vec,AIT.sq.vec/sum(AIT.sq.vec),dist)
names(out.list) <- c("Yhat","Distance Contributions",
```

```
  "Proportional Distance Contributions","Final Distance")

  return(out.list)
}
QFASA.const.eqn <- function(alpha,seal=seal.mat[i,],
prey.quantiles=prey.mat, gamma=gamma) {

  return(sum(alpha))
}

mean.geometric <- function(x) {

  # RETURNS GEOMETRIC MEAN

  # INPUT:
  #  x --> VECTOR

  D <- length(x)
  return(prod(x)^(1./D))
}




## Link function
btoa <- function(Covmat, coefbetas){
  mult <- Covmat%*%coefbetas
  alphs <- matrix(NA,nrow=nrow(Covmat), ncol=(ncol(coefbetas)+1))
  for(i in 1:nrow(mult)){ # number of preds
    for(j in 2:(ncol(mult)+1)){ # number of prey species
      alphs[i,j] <- exp(mult[i,(j-1)])
    }
    alphs[i,(1)] <- 1
```

```
  }
  alphs <- alphs/rowSums(alphs)
  return(alphs)
}



## Approximation of Betas

atob <- function(Xcov, alp){
  n <- nrow(Xcov)
  I <- ncol(alp)
  Xinv <- ginv(Xcov)
  S <- 1/alp[,1]
  almat <- alp[,-1]/alp[,1]
  almat <- log(almat)
  bet <- Xinv%*%almat
  return(bet)
}
```

Bibliography

Aitchison, J. (1986). *The statistical analysis of compositional data*. Chapman & Hall, Ltd.

Aitchison, J. (1992). On criteria for measures of compositional difference. *Mathematical Geology* **24**, 365–379.

Aitchison, J. & Bacon-Shone, J. (1999). Convex linear combinations of compositions. *Biometrika* **86**, 351–364.

Aitchison, J. M. (2005). A concise guide to compositional data analysis.

Arriola, A., Biuw, M., Walton, M., Moss, S. & Pomeroy, P. (2013). Selective blubber fatty acid mobilization in lactating gray seals (halichoerus grypus). *Physiological and Biochemical Zoology* **86**, 441–450.

Austin, D., Bowen, W., McMillan, J. & Boness, D. (2006). Stomach temperature telemetry reveals temporal patterns of foraging success in a free-ranging marine mammal. *Journal of Animal Ecology* **75**, 408–420.

Azzalini, A. & Valle, A. (1996). The multivariate skew-normal distribution. *Biometrika* **83**, 715–726.

Badjeck, M., Allison, E., Halls, A. & Dulvy, N. (2010). Impacts of climate variability and change on fishery-based livelihoods. *Marine policy* **34**, 375–383.

Beck, C., Bowen, W., McMillan, J. & Iverson, S. (2003a). Sex differences in diving at multiple temporal scales in a size-dimorphic capital breeder. *Journal of Animal Ecology* **72**, 979–993.

Beck, C., Bowen, W., McMillan, J. & Iverson, S. (2003b). Sex differences in the diving behaviour of a size-dimorphic capital breeder: the grey seal. *Animal Behaviour* **66**, 777–789.

Beck, C., Iverson, S., Bowen, W. & Blanchard, W. (2007). Sex differences in grey seal diet reflect seasonal variation in foraging behaviour and reproductive expenditure: evidence from quantitative fatty acid signature analysis. *Journal of Animal Ecology* **76**, 490–502.

Beckmann, C., Mitchell, J., Stone, D. & Huveneers, C. (2013). A controlled feeding experiment investigating the effects of a dietary switch on muscle and liver fatty acid profiles in port jackson sharks *Heterodontus portusjacksoni*. *Journal of Experimental Marine Biology and Ecology* **448**, 10–18.

Bowen, W. & Harrison, G. (1994). Offshore diet of grey seals halichoerus grypus near sable island, canada. *Marine ecology progress series. Oldendorf* **112**, 1–11.

Bowen, W., Lawson, J. & Beck, B. (1993). Seasonal and geographic variation in the species composition and size of prey consumed by grey seals (halichoerus grypus) on the scotian shelf. *Canadian Journal of Fisheries and Aquatic Sciences* **50**, 1768–1778.

Bowen, W., Oftedal, O. & Boness, D. (1992). Mass and energy transfer during lactation in a small phocid, the harbor seal (phoca vitulina). *Physiological Zoology* **65**, 844–866.

Brander, K. (2010). Impacts of climate change on fisheries. *Journal of Marine Systems* **79**, 389–402.

Breed, G., Bowen, W., McMillan, J. & Leonard, M. (2006). Sexual segregation of seasonal foraging habitats in a non-migratory marine mammal. *Proceedings of the Royal Society B: Biological Sciences* **273**, 2319–2326.

Bromaghin, J. (2015). Simulating realistic predator signatures in quantitative fatty acid signature analysis. *Ecological informatics* **30**, 68–71.

Bromaghin, J. (2017). qfasar: quantitative fatty acid signature analysis with r. *Methods in Ecology and Evolution* **8**, 1158–1162.

Bromaghin, J., Budge, S. & Thiemann, G. (2016). Should fatty acid signature proportions sum to 1 for diet estimation? *Ecological research* **31**, 597–606.

Bromaghin, J., Budge, S. & Thiemann, G. (2017a). Detect and exploit hidden structure in fatty acid signature data. *Ecosphere* **8**, e01896.

Bromaghin, J., Budge, S., Thiemann, G. & Rode, K. (2017b). Simultaneous estimation of diet composition and calibration coefficients with fatty acid signature data. *Ecology and Evolution* **7**, 6103–6113.

Bromaghin, J., Rode, K., Budge, S. & Thiemann, G. (2015). Distance measures and optimization spaces in quantitative fatty acid signature analysis. *Ecology and evolution* **5**, 1249–1262.

Budge, S., Iverson, S., Bowen, W. & Ackman, R. (2002). Among-and within-species variability in fatty acid signatures of marine fish and invertebrates on the scotian shelf, georges bank, and southern gulf of st. lawrence. *Canadian Journal of Fisheries and Aquatic Sciences* **59**, 886–898.

Budge, S., Iverson, S. & Koopman, H. (2006). Studying trophic ecology in marine ecosystems using fatty acids: a primer on analysis and interpretation. *Marine Mammal Science* **22**, 759–801.

Bühlmann, P. & Van De Geer, S. (2011). *Statistics for high-dimensional data: methods, theory and applications*. Springer Science & Business Media.

Cheung, W., Lam, V., Sarmiento, J., Kearney, K., Watson, R., Zeller, D. & Pauly, D. (2010). Large-scale redistribution of maximum fisheries catch potential in the global ocean under climate change. *Global Change Biology* **16**, 24–35.

Davison, A. & Hinkley, D. (1997). *Bootstrap methods and their application*, vol. 1. Cambridge university press.

Egozcue, J., Pawlowsky-Glahn, V., Mateu-Figueras, G. & Barcelo-Vidal, C. (2003). Isometric logratio transformations for compositional data analysis. *Mathematical Geology* **35**, 279–300.

Filzmoser, P., Hron, K., Reimann, C. & Garrett, R. (2009). Robust factor analysis for compositional data. *Computers & Geosciences* **35**, 1854–1861.

Finch, W. & Finch, M. H. (2017). Multivariate regression with small samples: A comparison of estimation methods. *Gen. Linear Model J.* **43**, 16–30.

Folch, J., Lees, M. & Stanley, G. (1957). A simple method for the isolation and purification of total lipides from animal tissues. *Journal of biological chemistry* **226**, 497–509.

Froese, R. & Kesner-Reyes, K. (2002). Impact of fishing on the abundance of marine species. ICES Council Meeting Report CM.

Gannes, L., O'Brien, D. & del Rio, C. (1997). Stable isotopes in animal ecology: assumptions, caveats, and a call for more laboratory experiments. *Ecology* **78**, 1271–1276.

Greenacre, M. (2011). Measuring subcompositional incoherence. *Mathematical Geosciences* **43**, 681–693.

Gueorguieva, R., Rosenheck, R. & Zelterman, D. (2008). Dirichlet component regression and its applications to psychiatric data. *Computational statistics & data analysis* **52**, 5344–5355.

Hobson, K. (1993). Trophic relationships among high arctic seabirds: insights from tissue-dependent stable-isotope models. *Marine Ecology-Progress Series* **95**, 7–7.

Hyslop, E. (1980). Stomach contents analysis?a review of methods and their application. *Journal of fish biology* **17**, 411–429.

Iverson, S., Field, C., Bowen, W. & Blanchard, W. (2004a). Quantitative fatty acid signature analysis: a new method of estimating predator diets. *Ecological Monographs* **74**, 211–235.

Iverson, S., Field, C., Bowen, W. & Blanchard, W. (2004b). Quantitative fatty acid signature analysis: A new method of estimating predator diets. *Ecological Monographs* **74**, 211–235.

Iverson, S., Frost, K. & Lowry, L. (1997). Fatty acid signatures reveal fine scale structure of foraging distribution of harbor seals and their prey in prince william sound, alaska. *Marine Ecology Progress Series* **151**, 255–271.

Iverson, S., Lang, S. & Cooper, M. (2001). Comparison of the bligh and dyer and folch methods for total lipid determination in a broad range of marine tissue. *Lipids* **36**, 1283–1287.

Kaunzinger, C. & Morin, P. (1998). Productivity controls food-chain properties in microbial communities. *Nature* **395**, 495.

Kelleher, K., Willmann, R. & Arnason, R. (2009). *The sunken billions: the economic justification for fisheries reform*. The World Bank.

Kirsch, P., Iverson, S. & Bowen, W. (2000). Effect of a low-fat diet on body composition and blubber fatty acids of captive juvenile harp seals (phoca groenlandica). *Physiological and Biochemical Zoology* **73**, 45–59.

Kristensen, K., Nielsen, A., Berg, C., Skaug, H. & Bell, B. (2016). Tmb: automatic differentiation and laplace approximation. *Journal of Statistical Software* .

Lancaster, H. (1965). The helmert matrices. *The American Mathematical Monthly* **72**, 4–12.

Lander, M., Harvey, J. & Gulland, F. (2003). Hematology and serum chemistry comparisons between free-ranging and rehabilitated harbor seal (phoca vitulina richardsi) pups. *Journal of Wildlife Diseases* **39**, 600–609.

Lang, S., Iverson, S. & Bowen, W. (2009). Repeatability in lactation performance and the consequences for maternal reproductive success in gray seals. *Ecology* **90**, 2513–2523.

Lang, S., Iverson, S. & Bowen, W. (2011). The influence of reproductive experience on milk energy output and lactation performance in the grey seal (halichoerus grypus). *PloS one* **6**.

Lidgard, D., Boness, D., Bowen, W. & McMillan, J. (2003). Diving behaviour during the breeding season in the terrestrially breeding male grey seal: implications for alternative mating tactics. *Canadian journal of zoology* **81**, 1025–1033.

Lidgard, D., Bowen, W. & Iverson, S. (2020). Sex-differences in fine-scale home-range use in an upper-trophic level marine predator. *Movement ecology* **8**, 1–16.

MacNeil, M., Graham, N., Cinner, J., Dulvy, N., Loring, P., Jennings, S., Polunin, N., Fisk, A. & McClanahan, T. (2010). Transitional states in marine fisheries: adapting to predicted global change. *Philosophical Transactions of the Royal Society B: Biological Sciences* **365**, 3753–3763.

Martín-Fernández, J., Barceló-Vidal, C., Pawlowsky-Glahn, V., Buccianti, A., Nardi, G. & Potenza, R. (1998). Measures of difference for compositional data and hierarchical clustering methods. In *Proceedings of IAMG*, vol. 98.

Martín-Fernández, J. & Thió-Henestrosa, S. (2006). Rounded zeros: some practical aspects for compositional data. *Geological Society, London, Special Publications* **264**, 191–201.

Martin-Fernandez, J. A., Palarea-Albaladejo, J. & Olea, R. A. (2011). Dealing with zeros. *Compositional data analysis: Theory and applications* , 43–58.

Mateu-Figueras, G., Pawlowsky-Glahn, V. & Barceló-Vidal, C. (2005). The additive logistic skew-normal distribution on the simplex. *Stochastic Environmental Research and Risk Assessment* **19**, 205–214.

McArdle, B. & Anderson, M. (2001). Fitting multivariate models to community data: a comment on distance-based redundancy analysis. *Ecology* **82**, 290–297.

Mellish, J., Iverson, S. & Bowen, W. (1999). Variation in milk production and lactation performance in grey seals and consequences for pup growth and weaning characteristics. *Physiological and Biochemical Zoology* **72**, 677–690.

Nordstrom, C., Wilson, L., Iverson, S. & Tollit, D. (2008). Evaluating quantitative fatty acid signature analysis (qfasa) using harbour seals phoca vitulina richardsi in captive feeding studies. *Marine Ecology Progress Series* **360**, 245–263.

Noren, S., Iverson, S. & Boness, D. (2005). Development of the blood and muscle oxygen stores in gray seals (halichoerus grypus): implications for juvenile diving capacity and the necessity of a terrestrial postweaning fast. *Physiological and Biochemical Zoology* **78**, 482–490.

Olive, D. (2016). Bootstrapping hypothesis tests and confidence regions. *preprint, see http://lagrange. math. siu. edu/Olive/ppvselboot. pdf* .

Palarea-Albaladejo, J. & Martín-Fernández, J. (2008). A modified em alr-algorithm for replacing rounded zeros in compositional data sets. *Computers & Geosciences* **34**, 902–917.

Palarea-Albaladejo, J. & Martin-Fernandez, J. (2013). Values below detection limit in compositional chemical data. *Analytica chimica acta* **764**, 32–43.

Pauly, D., Alder, J., Booth, S., Cheung, W., Christensen, V., Close, C., Sumaila, U., Swartz, W., Tavakolie, A., Watson, R. et al. (2008). Fisheries in large marine ecosystems: descriptions and diagnoses. *The UNEP large marine ecosystem report: a perspective on changing conditions in LMEs of the World?s Regional Seas. UNEP Regional Seas Reports and Studies* , 23–40.

Pawlowsky-Glahn, V. & Buccianti, A. (2011). *Compositional data analysis: Theory and applications*. John Wiley & Sons.

Pearson, K. (1897). Mathematical contributions to the theory of evolution.on a form of spurious correlation which may arise when indices are used in the measurement of organs. *Proceedings of the royal society of london* **60**, 489–498.

Petchey, O., McPhearson, P., Casey, T. & Morin, P. (1999). Environmental warming alters food-web structure and ecosystem function. *Nature* **402**, 69.

Royston, P. (1995). Remark as r94: A remark on algorithm as 181: The w-test for normality. *Journal of the Royal Statistical Society. Series C (Applied Statistics)* **44**, 547–551.

Signorell, A. (2019). *DescTools: Tools for Descriptive Statistics*. R package version 0.99.31.

Steeves, H., Mcmeans, B., Field, C., Stewart, C., Arts, M., Fisk, A., Lydersen, C., Kovacs, K. & Macneil, M. (2016). Non-parametric analysis of the spatio-temporal variability in the fatty-acid profiles among greenland sharks. *Journal of the Marine Biological Association of the United Kingdom* , 1–7.

Stewart, C. (2005). Inference on the diet of predators using fatty acid signatures .

Stewart, C. (2013). Zero-inflated beta distribution for modeling the proportions in quantitative fatty acid signature analysis. *Journal of Applied Statistics* **40**, 985–992.

Stewart, C. (2017). An approach to measure distance between compositional diet estimates containing essential zeros. *Journal of Applied Statistics* **44**, 1137–1152.

Stewart, C. & Field, C. (2011). Managing the essential zeros in quantitative fatty acid signature analysis. *Journal of Agricultural, Biological, and Environmental Statistics* **16**, 45–69.

Stewart, C., Iverson, S. & Field, C. (2014). Testing for a change in diet using fatty acid signatures. *Environmental and Ecological Statistics* , 1–18.

Thapanand, T., Jutagatee, T., Wongrat, P., Lekcholayut, T., Meksumpun, C., Janekitkarn, S., Rodloi, A., Moreau, J. & Wongrat, L. (2009). Trophic relationships and ecosystem characteristics in a newly-impounded man-made lake in thailand. *Fisheries management and ecology* **16**, 77–87.

Tsagris, M., Preston, S. & Wood, A. (2011). A data-based power transformation for compositional data. *Proceedings of the 4th Compositional Data Analysis Workshop, Girona, Spain* .

Tsagris, M. & Stewart, C. (2018a). A dirichlet regression model for compositional data with zeros. *Lobachevskii Journal of Mathematics* **39**, 398–412.

Tsagris, M. & Stewart, C. (2018b). A folded model for compositional data analysis. *arXiv preprint arXiv:1802.07330* .

van den Boogaart, K., Tolosana, R. & Bren, M. (2014). *compositions: Compositional Data Analysis*. R package version 1.40-1.

Vander Zanden, M., Casselman, J. & Rasmussen, J. (1999). Stable isotope evidence for the food web consequences of species invasions in lakes. *Nature* **401**, 464.

Wootton, J., Parker, M. & Power, M. (1996). Effects of disturbance on river food webs. *Science* **273**, 1558–1561.