NON-PARAMETRIC STATISTICAL TESTS FOR DIFFERENCES IN FATTY ACID COMPOSITION OF GREENLAND SHARKS

by

Holly Steeves

Submitted in partial fulfillment of the
requirements for the degree of
Master of Computer Science

at

Dalhousie University
Halifax, Nova Scotia
November 2013

Contents

List of Tables . . . . . **iv**

List of Figures . . . . . **v**

Abstract . . . . . **vi**

Glossary . . . . . **vii**

Chapter 1 Introduction **1**

1.1 Climate Change 1

1.2 Quantitative Fatty Acid Signature Analysis 2

1.3 MANOVA techniques 4

1.4 Real-Life Data 5

1.5 Thesis Outline 6

Chapter 2 Methods **8**

2.1 Compositional data 8

2.2 Issues with Compositional Data 9

2.3 Distances for Compositional Data 11

2.4 Problems with Zeros 14
2.4.1 Replacement of Zeros Method 14
2.4.2 Chi-Squared Distance 15

2.5 Non-Parametric MANOVA 17

Chapter 3 Simulation Study **22**

3.1 Pseudo-Predators 22

3.2 Computing the simulations 24

3.3 Simulation Results 25

Chapter 4 Results **32**

Chapter 5 Conclusions **40**

5.1 Summary 40
5.1.1 Simulations 41
5.1.2 Ecological Data 43

5.2 Future Research 44

Appendix A Chi-Squared Distance **46**

Appendix B Non-Parametric MANOVA **54**

Bibliography **62**

List of Tables

Table 3.1 True diets of the pseudo-seals used in the simulations 23

Table 3.2 Chi-squared distances between diets using $\gamma = 1/3$ 23

Table 3.3 Diet combinations used for the two-way simulations 25

Table 3.4 Point estimates and 99% binomial confidence intervals for the type I error simulation results for 1000 simulations for 3 significance levels using sample sizes of 15 and 45 and with both the chi-squared distance measure and Aitchison's distance measure 26

Table 3.5 Point estimates and 99% binomial confidence intervals for the type I error simulation results for 1000 simulations and 3 significance levels 29

List of Figures

Figure 3.1 Power of simulations using sample size $n = 15$ and using both Aitchison's and chi-squared distances for various significance levels and effect sizes 27

Figure 3.2 Power of simulations using sample size $n = 45$ and using both Aitchison's and chi-squared distances for various significance levels and effect sizes 28

Figure 3.3 Power results for the two-way simulations for a factor A effect using various significance levels 30

Figure 3.4 Power results for the two-way simulations for a factor B effect using various significance levels 31

Figure 4.1 Sample mean fatty acid signatures of Greenland sharks at both Svalbard and Cumberland Sound 33

Figure 4.2 Sample mean fatty acid signatures of Greenland sharks in Cumberland Sound collected during winter 2008-2009 36

Figure 4.3 Sample mean fatty acid signatures of Greenland sharks in Cumberland Sound collected during summer 2007-2009 37

Figure 4.4 Sample mean fatty acid signatures of Greenland sharks at both Svalbard and Cumberland Sound collected during summer 2008 38

Figure 4.5 Box plot of the relative differences between the mean fatty acid signatures of Greenland Sharks at Svalbard and Cumberland Sound collected during summer 2008 39

Abstract

Variations in predator diets is important in ecology to help us understand their top-down effects on the ecosystem. In predator diets, their fatty acid signatures reflect the proportions of prey consumed. Since fatty acid signatures are compositional and often longer than the sample size, a standard MANOVA test is unsuitable. Here, non-parametric MANOVA techniques are developed to test for differences in fatty acid signatures among locations, years, and seasons which infer differences in diets. Simulations show that the test has good power and appropriate type I error rates. The tests developed were applied to data on Greenland Sharks to test for differences in diets between individuals from Cumberland Sound, Canada, versus those from Svalbard, Norway and whether there is a yearly and/or seasonal effect on the diets. Diet compositions were found to vary between the locations, seasons and years, possibly caused by differing prey species distributions, migrations, and climate change.

Glossary

$AIT(\mathbf{X}_1, \mathbf{X}_2)$	Aitchison's distance between $\mathbf{X}_1$ and $\mathbf{X}_2$
$C(\mathbf{X})$	The closure operator which ensures that a subset sums to 1
$CS(\mathbf{X_1}, \mathbf{X_2})$	Chi-squared distance between $\mathbf{X_1}$ and $\mathbf{X_2}$
$ED(\mathbf{X}_1, \mathbf{X}_2)$	Euclidean distance between $\mathbf{X}_1$ and $\mathbf{X}_2$
G	Gower's centered matrix
$KL(\mathbf{X}_1, \mathbf{X}_2)$	Kullback-Leibler distance between $\mathbf{X}_1$ and $\mathbf{X}_2$
$L^d(\boldsymbol{\mu}, \boldsymbol{\Sigma})$	Additive logistic normal distribution with parameters $\boldsymbol{\mu}$, $\boldsymbol{\Sigma}$
$M^d(\boldsymbol{\mu}, \boldsymbol{\Sigma})$	Multiplicative logistic normal distribution with parameters $\boldsymbol{\mu}$, $\boldsymbol{\Sigma}$
$N^d(\boldsymbol{\mu}, \boldsymbol{\Sigma})$	Multivariate normal distribution with mean $\boldsymbol{\mu}$, and standard deviation $\boldsymbol{\Sigma}$
$SN^d(\boldsymbol{\mu}, \boldsymbol{\Sigma})$	Skew-normal distribution with mean $\boldsymbol{\mu}$, and standard deviation $\boldsymbol{\Sigma}$
S^d	The d-dimensional simplex
Y	Response variable in a linear model

$Z_{1-\alpha}$	The α^{th} percentile of the standard normal distribution
α	The significance level or targeted type I error rate of the test
β	A matrix of parameters in a linear model
δ_j	Imputation value used in the replacement methods
ϵ	Normal random variate representing error in a linear model
γ	Value by which a composition is power transformed in the chi-squared distance measure
$\mathbb{R}_+^D$	Positive D-dimensional real numbers
$\mathbf{X} = (x_1, x_2, ..., x_D)$	A composition of length D where composition means the elements are non-negative and sum to 1
$\mathbf{r} = (r_1, r_2, ..., r_D)$	Composition which has had its zeros replaced using an imputation method
$d(\mathbf{X}_1, \mathbf{X}_2)$	Distance between $\mathbf{X}_1$ and $\mathbf{X}_2$ calculated using any metric or semi-metric distance measure
$f(\mathbf{X})$	Some function of $\mathbf{X}$
$g(\mathbf{X})$	Geometric mean of the vector $\mathbf{X}$

$tr(\mathbf{A})$	Trace of the matrix $\mathbf{A}$

Chapter 1

Introduction

The role that animals play in marine ecosystems is among the most important aspects of marine ecology. It allows ecologists to understand the structuring of marine ecosystems and the interactions between species. It is also useful in predicting possible consequences of fisheries, pollution, invasive species, or climate change on certain species or the ecosystem as a whole. Particularly, we are interested in how climate change is affecting species distributions and detecting any resultant change in animal diet (MacNeil et al. (2010)).

1.1 Climate Change

Climate change is dramatically altering ecosystems around the world. Among those most critically affected are coral reefs (Graham et al. (2008)) and the polar regions (MacNeil et al. (2010)). Coral reefs are threatened by climate change because corals are highly sensitive to changes in temperature. This loss of live coral leads to dramatic changes in the reef ecosystem including loss of reef structure, biodiversity, and the abundance of fish living on them. In contrast, polar regions are also threatened by global change due to warming waters and continuing loss of seasonal and multi-year ice (MacNeil et al. (2012)).

The Arctic Ocean ecosystem has already been strongly affected by climate change, as warming temperatures have caused a huge loss of both seasonal and multi-year sea ice that may soon lead to ice free summers (Overpeck et al. (2005)). This has a significant effect on all Arctic life. Loss of seasonal sea ice reduces algal growth that forms underneath its surface and is a primary source of production for the dominant benthic food-web. This loss of production is expected to be counteracted by increasing primary production as new ice free regions will allow more light below the surface that, combined with increased freshwater input, will increase productivity in the water column (Brander (2010)). This is thought to have dramatic effects on the distribution of

species within these areas. As water temperatures warm, marine species are expected to distribute themselves towards deeper water and higher latitudes (Cheung et al. (2010)). However, it is not known how resident diets might change as new species move into the Arctic. Such changes may have a large effect on the diets of Arctic predators as the availability and variety of prey species is altered. However, a key knowledge gap exists in understanding contemporary predator diets.

Baseline dietary information for predators is important for understanding the consequences of a changing prey base in the Arctic due to climate change. While a few diets can be estimated from direct observation, most marine predators feed below the surface. In such cases, stomach contents have traditionally been used to identify diet compositions of predators but this approach can be problematic (Beckmann et al. (2013)). One issue is that animals often need to be euthanized to obtain stomach contents. Killing animals is not ideal, especially for endangered or protected species, so a less intrusive method is desired. Secondly, there is a bias towards digestion-resistant hard parts because soft tissues are not persistent in the stomach and some soft-bodied prey species in the diet may therefore often not be found in stomach contents (Iverson et al. (2004)). Finally, the stomach content method is only indicative of the most recent meal, and not the long term diet composition, which is of interest.

Recently, a new method using fatty acids has been developed to identify the diet composition of predators. Fatty acids are the fundamental components of lipids that do not degrade during digestion. These lipids are fat that has been ingested and used to store energy (Iverson et al. (2004)). Some of these fatty acids are absorbed into the fat stores, or adipose tissues with little modification. Fatty acids can be divided into dietary and non-dietary components and used to estimate the diets of predators based on the fatty acids of the prey (Iverson et al. (2004)). This method of estimating the diets based on fatty acids is called quantitative fatty acid signature analysis or QFASA.

1.2 Quantitative Fatty Acid Signature Analysis

QFASA is a technique used to estimate the diet composition of a predator using fatty acid signatures of the predators and prey. A fatty acid signature is a vector that is defined to be the quantitative distribution of all the fatty acids present in an

individual (Iverson et al. (2004)). It is a specific type of vector called a composition (see section 2.1) where the elements are all non-negative and sum to 1. In order to use QFASA, a sample of fatty acid signatures of each possible prey type must be collected to form a prey database. For each prey species, a summary fatty acid signature is calculated, such as a mean. Weights are allocated to the summary fatty acid signature of each prey type such that the statistical distance from this weighted mixture to the fatty acid signature of the predator is minimized. These weights are also compositions as their elements are non-negative and sum to 1.

The selection of the distance measure to be used varies but the Kulback-Liebler distance was originally used with QFASA which is defined in section 2.3 (Iverson et al. (2004)). Another suggestion for the distance measure is Aitchison's distance described in section 2.3 which is preferred over the Kulback-Leibler distance because it satisfies subcompositional coherence (see section 2.3). The weights selected represent the estimated proportion of each prey type in the predator's diet composition.

Recently, calibration factors and fat content measurements have been added into fatty acid calculations to improve accuracy. Calibration factors are used because for certain fatty acids, the quantity in the predator may always be higher or lower than the original fatty acid quantity in the prey. These factors are determined by captive feeding experiments where the predators are fed a consistent diet long-term. Fat content is used because certain prey have a higher fat content and therefore contribute more to the fatty acid signature of the predator (Iverson et al. (2004)). These values are collected for each prey species with the data.

Often in ecology, interest lies in whether or not a difference in diets exists spatially or temporally in a species. If QFASA were utilized in this case, a large prey database and their fatty acid signatures would be required which is logistically difficult, particularly in polar environments. Stewart et al. ("In Review") argues that if the predators sampled are from the same region and season, and their fatty acid signatures are found to be significantly different, it implies that the diets are significantly different as well. In settings without a prey database, we are limited to testing for changes in fatty acid signatures. For this research, the predators sampled are from different areas measured in summer and winter over three years. In this situation, if the fatty acid signatures are found to be significantly different, it could be due to a

difference in diets but it is also possible that it is due to the same diet where the prey has different fatty acid signatures in the different regions, seasons, or years. However, we have no reason to believe that the fatty acid signatures of the prey are different so we will attribute a difference in fatty acid signatures to a difference in diets.

1.3 MANOVA techniques

Multivariate Analysis of Variance or MANOVA is a method used to test for differences in the means of different treatment groups. In order to apply this methodology to testing for differences in fatty acid signatures, certain modifications are required due to restrictions associated with fatty acids. Because all of the elements represent proportions, and are therefore non-negative and sum to one, they are classified as compositional data (Aitchison (1986)). Compositional data is challenging because of these restrictions, so special care is required when applying statistical methods. Aitchison (1986) proposes using log-ratio transformations on the data. If the transformed data satisfies the assumption of multivariate normality, then we could use a standard MANOVA technique on the transformed data. The first problem that arises with this is that log-ratio transformations cannot be used on data containing zeros. To solve this, a multiplicative replacement method is used to replace these zeros as if they were missing values in the data (Martín-Fernández & Thió-Henestrosa (2006)). Once the zeros are replaced with positive entries, the transformations can be performed and a standard MANOVA technique can be used. The second problem that arises is that often with fatty acid signatures, the number of fatty acids is much larger than the sample size. This creates a problem because a MANOVA can only be used when the number of dimensions is less than the sample size. For this issue, new techniques are developed.

McArdle & Anderson (2001), develop a non-parametric MANOVA technique that does not rely on the assumption that the log-ratio transformed data is multivariate normal, nor that the sample size is larger than the number of dimensions. Instead of using the original data in the calculation, they develop a distance matrix that represents the pairwise distances between each combination of fatty acid signatures among individuals. A modification to the usual breakdown of sum of squares is performed that is based on traces of matrices. This allows the data matrices in the

sum of squares formulas to be replaced with the analogous Gower's centered distance matrix. This distance matrix can be calculated using any distance measure or semi-metric distance measure where semi-metric means that the triangle inequality may not hold true for that distance. This allows us to use a replacement method for the zeros, Aitchison's distance measure, or an alternative distance measure like the chi-squared distance which is capable of handling zeros. When the Euclidean distance measure is used and the assumption of normality is valid, this method will give the same statistic value as the standard MANOVA statistic. This should be a useful technique in ecology as many ecological data sets do not follow the strict assumptions of a standard MANOVA.

1.4 Real-Life Data

Significant changes in the Arctic due to climate changes (described in section 1.1) are affecting the dynamics of the Arctic marine food web. For this reason, trophic dynamics in the Arctic are of interest in order to understand potential consequences of climate change within the ecosystem (MacNeil et al. (2012)). Apex predators comprise a major component in ecosystem dynamics as they are thought to exert top-down control on marine food webs. In the Arctic, one such predator is the Greenland shark, which has not been well studied to date. As a potentially abundant apex predator, the Greenland shark may play a critical role in Arctic ecosystems.

Greenland sharks near Svalbard, Norway and Cumberland Sound, Canada are thought to comprise two distinct populations in the Arctic Ocean. Svalbard is located in the middle of the Arctic Ocean, east of Greenland and north of Europe (78° N, 16° E). Cumberland sound is off of the Davis Strait to the west of Greenland (65° 13′ N 65° 45′ W). It is hypothesized that their spatial separation may lead to differing diets at the two locations that might change substantially with a predicted inflow of new species northward due to climate change. In order to test for hypothesized differences in diet, fatty acid signatures of Greenland Sharks were measured during winter and summer for several years and at the two locations.

We are interested in finding a test compatible with our data set to test whether or not the sharks have different diets at the two locations. Because a prey database is not available, QFASA cannot be used to obtain diet estimates for comparison. We can

however test for significant differences in fatty acid signatures and infer differences in diet (Stewart et al. ("In Review")). The dimension of the fatty acid signature is larger than the number of sharks sampled so we cannot use a standard MANOVA. Instead, we use the non-parametric MANOVA technique proposed by McArdle & Anderson (2001) to test for differences in diets at the two locations. We are also interested in whether there is a seasonal effect and/or yearly effect on the diets. To test this, an unbalanced three-way non-parametric MANOVA is used based on a combination of a partial F-test technique and the non-parametric sum of squares presented in McArdle & Anderson (2001).

1.5 Thesis Outline

The methods and techniques used in the simulations to test the non-parametric MANOVA approach and to test for differences in the Greenland Shark data are introduced and explained in detail in Chapter 2. To begin this chapter, compositional data is defined as well as several concepts associated with compositional data including the simplex. This is presented in section 2.1. The following section describes previous methods and models used on compositional data such as log-ratio techniques and the Dirichlet model, as well as challenges and problems associated with using these methods. Section 2.3 presents different types of distances that can be used including Euclidean distance, Aitchison's distance, and Kulback-Leibler distance, and the restrictions on each one. One of these restrictions is that many distances do not allow zero elements in the composition, which is a common occurrence in ecological data. This issue is introduced in section 2.4 and some possible solutions to this issue are presented in section 2.4.1, including an imputation method that treats zeros as missing values in order to replace them. Instead of replacing the zeros, a measure could be used that is capable of handling zeros, such as the chi-squared distance defined in section 2.4.2. Both these solutions to zeros are useful when applying compositional techniques to data with zeros. A non-parametric technique that requires these solutions is explained in section 2.5. This section describes both a one-way and two-way non-parametric MANOVA method that allows the sample size to be larger than the number of dimensions which is a useful feature when dealing with fatty acid data.

Next, Chapter 3 describes simulations that were conducted in order to assess the

type I error and power of the one-way and two-way non-parametric MANOVA as well as the results obtained. It describes in detail how simulations were constructed with realistic predator diets based on seals on the East coast of Canada and using both Aitchison's and chi-squared distance measures described in sections 2.3 and 2.4.2. In Chapter 4, the methods described in Chapter 2 are applied to real ecological data. The data set contains fatty acid signatures of Greenland sharks taken both in summer and winter from 2007 until 2009 collected near Svalbard and Cumberland Sound. Based on the results obtained from this data and the results from the simulations, conclusions are formed in Chapter 5 as well as discussion of future research.

Chapter 2

Methods

2.1 Compositional data

Aitchison (1986) defines a composition to be a vector $\mathbf{X} = (x_1, x_2, ..., x_D)$ that has non-negative elements and satisfies the unit-sum constraint. The unit-sum constraint refers to the elements of the vector $\mathbf{X}$ summing to one:

$$x_1 + x_2 + ... + x_D = 1$$

Compositions represent proportions or percentages and are common in many different fields. In our application, the data represent percentages of each type of fatty acid present in a tissue sample from an individual animal. A subcomposition is defined to be a closed subset of elements of a composition. Closed means that a closure operator has been applied to the subset of elements, which ensures that the subset sums to 1. This operation is defined as follows (Pawlowsky-Glahn & Buccianti (2011)):

Definition 1. *The closure operator* $C(\boldsymbol{X})$ *is*

$$C(\boldsymbol{X}) = \left(\frac{x_1}{\sum_{i=1}^{D} x_i}, \frac{x_2}{\sum_{i=1}^{D} x_i}, ..., \frac{x_D}{\sum_{i=1}^{D} x_i} \right)$$

Compositional data is defined on a d-dimensional simplex where $d = D - 1$ which Aitchison (1986) defined as the set S^d such that:

$$S^d = \{(x_1, ..., x_d) : x_1 > 0, ..., x_d > 0; x_1 + ... + x_d < 1\}$$

Because of the constraints that compositional data present, several difficulties arise relating to the assumptions of parametric modelling, spurious correlation, and negative bias difficulty.

2.2 Issues with Compositional Data

When typical statistical analyses are performed on compositional data, it can often lead to irrelevant or incorrect results due to the strict constraints on the data. The most common misleading result is spurious or false correlation. Spurious correlation says that if we let $x_1 = f(w_1, w_3)$ and $x_2 = f(w_2, w_3)$, where f is some function and there is no pairwise correlation between w_1, w_2 and w_3, there will be correlation between x_1 and x_2 due to the function itself. This applies to compositions because, for example, given the vector (w_1, w_2), with $w_3 = w_1 + w_2$, the vector can be converted into a composition by applying the closure operator defined in Definition 1. This will yield the composition $(x_1, x_2) = (f(w_1, w_3), f(w_2, w_3))$ where $f(a, b) = \frac{a}{b}$. This function of ratios leads to false correlation, or spurious correlation between the elements of the composition x_1 and x_2.

Other compositional data problems include restrictions to non-negative entries and the unit-sum constraint, both of which make parametric modelling difficult. For example, consider a linear model, $Y = \beta X + \epsilon$, where ϵ is a normal random variate. The predicted values Y are not restricted in this model. If Y is a d-dimensional vector, the values of Y are free to vary through d-dimensional space. However, when dealing with compositional data, Y is restricted to the d-dimensional simplex. Aitchison (1986) presents several different models that work well with compositional data, including the Dirichlet distribution, the additive logistic normal distribution, and the multiplicative logistic normal distribution.

Definition 2. *The Dirichlet distribution defined on the d-dimensional simplex S^d with parameter $\boldsymbol{\alpha} \epsilon \mathbb{R}_+^D$ has density function:*

$$\frac{\Gamma(\alpha_1 + ... + \alpha_D)}{\Gamma(\alpha_1) \cdots \Gamma(\alpha_D)} x_1^{\alpha_1 - 1} \cdots x_d^{\alpha_d - 1} (1 - x_1 - ... - x_d)^{\alpha_D - 1} \qquad (\boldsymbol{X} \epsilon S^d)$$

where d = D - 1.

Definition 3. *A D-part composition $\boldsymbol{X}$ is defined to be additive logistic normal $L^d(\boldsymbol{\mu}, \boldsymbol{\Sigma})$ when $\boldsymbol{Y} = log(\boldsymbol{X_{-D}}/x_D)$, where $\boldsymbol{X_{-D}}$ is the vector $\boldsymbol{X}$ without part D, follows a normal distribution $N^d(\boldsymbol{\mu}, \boldsymbol{\Sigma})$.*

Definition 4. *A D-part composition* $\boldsymbol{X}$ *is defined to be multiplicative logistic normal* $M^d(\boldsymbol{\mu}, \boldsymbol{\Sigma})$ *when* $\boldsymbol{Y^{(d)}}$*, defined by*

$$y_i = log\{x_i/(1 - x_1 - ... - x_i)\} \qquad (i = 1, ..., d)$$

follows a normal distribution $N^d(\boldsymbol{\mu}, \boldsymbol{\Sigma})$ *.*

The Dirichlet distribution has some important limitations including that the composition must have subcompositional independence, meaning that the set of all subcompositions must be independent for all partitions of the original composition (Pawlowsky-Glahn & Buccianti (2011)). If a composition satisfies subcompositional independence, each ratio of two components $\frac{x_i}{x_j}$ is independent of any other ratio of two components $\frac{x_m}{x_n}$ which is very unlikely to be satisfied in many kinds of compositional data (Aitchison (1986)). The alternative distributions are less restrictive and transform the compositional data approximately to normality through logarithmic transformations. Due to the assumed log-normality of the data, if this assumption does not hold, we cannot use the distributions to model the data as results will be invalid. If we find that the data is not log-normal, but follows an asymmetrical distribution, there are alternative distributions that can be used. Logarithmic transformations can also be used to transform these compositions to a skew-normal distribution, called the additive logistic skew-normal distribution, which allows for modelling of data that does not follow a symmetrical distribution (Mateu-Figueras et al. (2005)).

Definition 5. *A D-part composition* $\boldsymbol{X}$ *is defined to have an additive logistic skew-normal distribution when* $\boldsymbol{Y} = log(\boldsymbol{X_{\text{-}D}}/x_D)$*, where* $\boldsymbol{X_{\text{-}D}}$ *is the vector* $\boldsymbol{X}$ *without part D, has a skew-normal distribution* $SN^d(\boldsymbol{\mu}, \boldsymbol{\Sigma})$ *.*

We can similarly define a multiplicative logistic skew-normal distribution (Stewart & Field (2011)).

Another difficulty that arises with compositional data is what Aitchison (1986) refers to as negative bias difficulty. This is a problem connecting the variances and covariances of the composition. It arises due to the equality shown below, where the first line holds true from basic principals of covariances.

$$\begin{aligned} Var[x_1] + Cov[x_1, x_2] + ... + Cov[x_1, x_D] &= Cov[x_1, x_1 + x_2 + ... + x_D] \\ &= Cov[x_1, 1] \\ &= 0 \end{aligned}$$

If we rearrange this, we get that $Cov[x_1, x_2] + ... + Cov[x_1, x_D] = -Var[x_1]$. This means that at least one of the covariances between the first element and some other element is negative. This unduly restricts the correlations, so that they are not free to take on any value between -1 and 1 which could cause problems for interpreting the correlation matrix.

2.3 Distances for Compositional Data

The difficulties compositional data present require an appropriate measure of distance between compositions, the most commonly used measure being Euclidean distance. If we let $\mathbf{X}_1$ and $\mathbf{X}_2$ be vectors, then the Euclidean distance between $\mathbf{X}_1$ and $\mathbf{X}_2$ is defined as follows:

$$ED(\mathbf{X}_1, \mathbf{X}_2) = \sqrt{\sum_{j=1}^{D} (x_{1j} - x_{2j})^2}$$

where D is the length of the vectors. Aitchison (1986) proposed an alternative to Euclidean distance that was designed specifically for compositional data. Aitchison's distance measure is defined as:

$$AIT(\mathbf{X_1}, \mathbf{X_2}) = \sum_{j=1}^{D} \{log[x_{1j}/g(\mathbf{X_1})] - log[x_{2j}/g(\mathbf{X_2})]\}^2 \tag{2.1}$$

where $g()$ represents the geometric mean computed by $g(\mathbf{X}) = (x_1 \cdot x_2 \cdots x_D)^{1/D}$. Another distance measure used on compositional data, and specifically in QFASA, is the Kulback-Leibler distance, defined as:

$$KL(\mathbf{X_1}, \mathbf{X_2}) = \sum_{j=1}^{D} (x_{1j} - x_{2j}) log(\frac{x_{1j}}{x_{2j}})$$

An important property for distances in compositional data is subcompositional coherence. Coherence is important because it justifies the use of ratios of parts instead of the original elements. Essentially, it means that the same conclusions and relationships can be found from studying a subcomposition as when studying the full composition. Subcompositional coherence is defined as follows (Pawlowsky-Glahn & Buccianti (2011)).

Definition 6. *A distance for compositional data is said to have subcompositional coherence if the following are satisfied:*

- *Scale invariance (definition 7)*
- *Subcompositional dominance (definition 8)*

Definition 7. *A distance is said to satisfy scale invariance if two compositions are scaled by a constant and the distance remains the same. That is, if* $\boldsymbol{X}_1$ *and* $\boldsymbol{X}_2$ *are compositions, and they are scaled by a constant* c*, then the distance* d *is said to satisfy scale invariance if:*

$$d(c\boldsymbol{X}_1, c\boldsymbol{X}_2) = d(\boldsymbol{X}_1, \boldsymbol{X}_2)$$

Definition 8. *A distance measure is said to satisfy subcompositional dominance when the distance obtained by comparing the full compositions is greater than or equal to the distance obtained by comparing the corresponding subcompositions. Let* d *be some distance measure,* $\boldsymbol{X}_1$ *and* $\boldsymbol{X}_2$ *be compositions, and* $\boldsymbol{X}_{s1}$ *and* $\boldsymbol{X}_{s2}$ *be their subcompositions respectively, then* d *is said to be subcompositionally dominant if*

$$d(\boldsymbol{X}_1, \boldsymbol{X}_2) \geq d(\boldsymbol{X}_{s1}, \boldsymbol{X}_{s2})$$

for all possible subcompositions $\boldsymbol{X}_{s1}$ *and* $\boldsymbol{X}_{s2}$*.*

Aitchison (1986) defines a subcomposition by the projection of the original composition onto a simplex. Therefore, subcompositional dominance is an important feature because it says that the subcomposition is behaving like a projection of the composition. When both scale invariance and subcompositional dominance are satisfied, the distance satisfies subcompositional coherence.

It can be shown that neither the Euclidean distance nor the KL distance satisfy subcompositional coherence. Take for example two compositions $\mathbf{X}_1 = (0.4, 0.4, 0.2)$ and $\mathbf{X}_2 = (0.7, 0.1, 0.2)$ and consider the subcomposition of $\mathbf{X}_1$, $\mathbf{X_{s1}} = (0.5, 0.5)$ and the subcomposition of $\mathbf{X}_2$, $\mathbf{X_{s2}} = (0.875, 0.125)$ found by taking the first two elements of each composition. If we compute the Euclidean distance between the two compositions, we get $ED(\mathbf{X}_1, \mathbf{X}_2) = 0.4243$, but the Euclidean distance between the two corresponding subcompositions is $ED(\mathbf{X_{s1}}, \mathbf{X_{s2}}) = 0.5303$. Similarly, the Kulback-Leibler distance between the compositions is $KL(\mathbf{X}_1, \mathbf{X}_2) = 0.2535$ but the Kulback-Leibler distance between the two corresponding subcompositions is $KL(\mathbf{X_{s1}}, \mathbf{X_{s2}}) = 0.3169$. This tells us that both the Euclidean distance and the Kulback-Leibler distance fail to satisfy subcompositional dominance.

It can also be shown that neither distance measure satisfies scale invariance. Take for example the vectors $\mathbf{Y}_1 = (1, 1)$ and $\mathbf{Y}_2 = (1, 2)$ and the corresponding vectors when scaled by $c = 2$, $\mathbf{Y_{c1}} = (2, 2)$ and $\mathbf{Y_{c2}} = (2, 4)$. The Euclidean distance between the original vectors is $ED(\mathbf{Y}_1, \mathbf{Y}_2) = 1$, but the Euclidean distance between the scaled vectors is $ED(\mathbf{Y_{c1}}, \mathbf{Y_{c2}}) = 2$. Similarly, the Kulback-Leibler distance between the original vectors is $KL(\mathbf{Y}_1, \mathbf{Y}_2) = 0.3010$ but the Kulback-Leibler distance between the scaled vectors is $KL(\mathbf{Y_{c1}}, \mathbf{Y_{c2}}) = 0.6020$. This tells us that these distances do not satisfy scale invariance. Therefore neither the Euclidean nor Kulback-Leibler distance measures are appropriate for compositional data (Pawlowsky-Glahn & Buccianti (2011)). However, if we compute Aitchison's distance between both $\mathbf{Y}_1$ and $\mathbf{Y}_2$, and $\mathbf{Y_{c1}}$ and $\mathbf{Y_{c2}}$, we get 0.0453 for both. This makes sense since when the vectors are closed, they are the same compositions. It can be shown that this is true for all possible vectors making Aitchison's distance scale invariant. Aitchison's distance therefore satisfies both features and so it is subcompositionally coherent. The proof that Aitchison's distance satisfies these properties can be found in (Aitchison (1992)). The distances that have been shown to have subcompositional coherence have been based on ratios (Martín-Fernández et al. (1998)). Aitchison's distance works well with compositional data, however a problem arises when the data contain zeros.

2.4 Problems with Zeros

Most methods and distances for compositional data involve logarithms and ratios, both of which lead to problems when there are zeros involved. This is a common problem because compositional data frequently have zero elements, including fatty acid profile data. There are three types of zeros (Pawlowsky-Glahn & Buccianti (2011)): essential zeros, count zeros, and rounded zeros. Essential zeros, also known as structural or absolute zeros, are zeros that represent the true absence of that element in the composition. Count zeros are similar to essential zeros except they occur in discrete compositional data or count data whereas essential zeros are found in continuous data. Rounded zeros are zeros that could be treated as a missing value because the true value of the element was below the rounding threshold and the element rounded to zero. Rounded zeros are the most common in compositional data and have had the most development for techniques to deal with them (Martín-Fernández & Thió-Henestrosa (2006), Pawlowsky-Glahn & Buccianti (2011), Aitchison (1986)).

2.4.1 Replacement of Zeros Method

Rounded zeros can be considered missing values in the data that can be dealt with by way of imputation, as proposed by Martín-Fernández & Thió-Henestrosa (2006). Imputation is a replacement strategy to fill the missing values with specific quantities. Martín-Fernández & Thió-Henestrosa (2006) describe three different replacement methods: additive replacement, simple replacement, and multiplicative replacement. They discovered that the multiplicative strategy tends to give the best results and so it is the approach we follow here. Consider δ_j to be the imputation value for part x_j which is a value that will be used in the replacement formulae. Pawlowsky-Glahn & Buccianti (2011) suggest that this imputation value be equal to 65% of the rounding threshold of part x_j. For instance, if the data is measured to 0.1 units, the rounding threshold would be 0.05 units since below 0.05 gets rounded to zero, and above gets rounded to 0.1. Then the imputation value would be 65% of 0.05 which is 0.032. We will use this imputation method to select the δ_j, but we recognize that other approaches are possible. The formulae for the replaced composition $\mathbf{r} = (r_1, r_2, ..., r_D)$ are as follows:

The Multiplicative Replacement Method

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

The Additive Replacement Method

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

The Simple Replacement Method

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

where $\mathbf{X}$ is a D-part composition with Z rounded zeros and c is the sum-constraint, usually 1 or 100%. The Simple Replacement method is the easiest method to understand as each zero is replaced by the imputation value δ_j and then the composition is reclosed. The Additive Replacement method is called "additive" because as every zero is replaced, every non-zero value is reduced by the amount $\frac{Z+1}{D^2} \left(\sum_{k|x_k=0} \delta_k\right)$ and the Multiplicative Replacement method is called "multiplicative" because every non-zero element is reduced by a factor of $1 - \frac{\sum_{k|x_k=0} \delta_k}{c}$. In the next subsection, we consider a measure that can handle zeros.

2.4.2 Chi-Squared Distance

For measuring the distance between samples of compositional data, Stewart et al. ("In Review") proposed a chi-squared distance measure based on the work of Greenacre (2011). The development of this distance was motivated by the desire for a statistical distance that satisfies subcompositional coherence but allows for handling zeros without having to modify or replace them. Chi-squared distance maintains scale invariance, does not require zeros to be changed, and the value γ is selected so that the chi-squared distance gets very close to subcompositional coherence. The data

is power transformed by this value γ and reclosed using the closure operation from Definition 1, where power transformed means that the vector was raised to the power γ.

Definition 9. *The chi-squared distance measure is defined as follows:*

$$CS(\boldsymbol{X_1}, \boldsymbol{X_2}) = \frac{1}{\gamma}\sqrt{2D}\left(\sum_{j=1}^{D}\frac{(Z_{1j} - Z_{2j})^2}{c_j}\right)^{1/2} \tag{2.2}$$

where $\boldsymbol{Z_1}$ and $\boldsymbol{Z_2}$ are the reclosed power transformations by γ of $\boldsymbol{X_1}$ and $\boldsymbol{X_2}$ respectively, $c_j = Z_{1j} + Z_{2j}$ and D is the length of the compositions. When Z_{1j} and Z_{2j} are both zero, the fraction is set to zero. When γ goes to zero, the chi-squared distance measure becomes approximately the square root of Aitchisons distance. The process of proving this property is explained in Greenacre (2011).

The power of the transformation γ is specific to the data and needs to be computed before calculating the distance measure. The calculation of γ is a pitfall for chi-squared distance as it requires samples of two populations (as opposed to simply two values) and is slow to calculate due to computations of distances required.

For two samples of compositional data that contain zeros, the process for finding the power of the transformation γ begins by calculating the average stress for each γ in a sequence of values such as $\frac{1}{1}$, $\frac{1}{2}$, $\frac{1}{3}$, ... Greenacre (2011) defines stress as follows:

$$stress = \sqrt{\sum\sum_{i<j}\frac{(\mathbf{D}_{ij} - \boldsymbol{\delta}_{ij})^2}{\sum\sum_{i<j}\mathbf{D}_{ij}^2}}$$

where $\mathbf{D}$ is the matrix of pairwise distances between the full compositions, and $\boldsymbol{\delta}$ is the matrix of pairwise distances between the corresponding two-part subcompositions. This definition for stress cannot be used with our application, as Greenacre was looking at distances between components, not compositions. Therefore, Stewart et al. ("In Review") propose a different formula for calculating stress that represents a measure of subcompositional incoherence. It is defined as:

$$stress = 1 - \frac{\#\{\mathbf{D}_{ij} > \boldsymbol{\delta}_{ij}\}}{n_1 n_2}$$

where n_1 and n_2 are the number of compositions or sample size in group 1 and 2

respectively. To compute the average stress for a given γ, the compositions are power transformed with γ then reclosed using Definition 1. All pairwise chi-square distances are calculated between the two transformed compositional data sets and then are compared to pairwise distances between two-part subcompositions using the stress formula from Stewart et al. ("In Review"). We are using two-part subcompositions because Greenacre (2011) found that they were those most affected by subcompositional incoherence. Average stress is simply the average of the stress values over all possible two-part subcompositions. This is repeated over the sequence of γ values and the γ with the lowest average stress value is selected. However, when there are no zeros (as in the simulations below), the stress value will continue to approach zero as γ decreases because chi-squared distance approaches the square root of Aitchison's distance which is subcompositionally dominant. The proof for this is based on the Box-Cox transformation (Greenacre (2010), Pawlowsky-Glahn & Buccianti (2011)). Because of this, if zeros are not present in the compositions, we select γ when the average stress is very close to zero. In order to obtain γ for three groups, the same approach is taken only all possible pairwise distances between the groups must be used. This γ is used to calculate the chi-squared distance described in section 2.4.2. This distance measure can be used in statistical methods such as the non-parametric MANOVA described in section 2.5 instead of using the replacement method to handle zeros.

2.5 Non-Parametric MANOVA

Aitchison (1986) proposes the possibility of completing a MANOVA on compositions by transforming the compositions using log-ratios. This method is 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 in the case with most fatty acid profiles. Also,

as discussed, Euclidean distances are not subcompositionally coherent, making them inappropriate measures for compositions.

McArdle & Anderson (2001) proposed a non-parametric MANOVA test that is appropriate for compositional data. With this test, 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. It is based on the theory behind a parametric MANOVA and the transposition of matrices.

To begin, we will describe the theory behind a standard multivariate MANOVA. A parametric MANOVA requires $\mathbf{Y}$, a sample of n units in p variables, as well as a nxp model matrix or design matrix $\mathbf{X}$. ANOVA exists to test the hypothesis that the model parameters have no effect. That is, $H_o : \boldsymbol{\beta} = \mathbf{0}$ in the equality $\mathbf{Y} = \mathbf{X}\boldsymbol{\beta} + \boldsymbol{\epsilon}$. It can be shown that the least squares solution for $\boldsymbol{\beta}$ is given by $\mathbf{B} = (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{Y}$. If we let $\hat{\mathbf{Y}}$ be the predicted values, then $\hat{\mathbf{Y}} = \mathbf{XB} = \mathbf{X}(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{Y}$ and we can define the hat matrix as $\mathbf{H} = \mathbf{X}(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'$ and therefore $\hat{\mathbf{Y}} = \mathbf{HY}$. The residuals for these predicted values will then be given by $\mathbf{R} = \mathbf{Y} - \hat{\mathbf{Y}} = \mathbf{Y} - \mathbf{HY} = (\mathbf{1} - \mathbf{H})\mathbf{Y}$. Typically, the total sum of squares, SST, is given by $SST = tr(\mathbf{Y}'\mathbf{Y})$, but if we expand $\mathbf{Y}'\mathbf{Y}$ in terms of the predicted values and the residuals, we get:

$$\begin{aligned}
\mathbf{Y}'\mathbf{Y} &= (\hat{\mathbf{Y}} + \mathbf{R})'(\hat{\mathbf{Y}} + \mathbf{R}) \\
&= (\hat{\mathbf{Y}}' + \mathbf{R}')(\hat{\mathbf{Y}} + \mathbf{R}) \\
&= \hat{\mathbf{Y}}'\hat{\mathbf{Y}} + \hat{\mathbf{Y}}'\mathbf{R} + \mathbf{R}'\hat{\mathbf{Y}} + \mathbf{R}'\mathbf{R}
\end{aligned}$$

Since $\hat{\mathbf{Y}}$ and $\mathbf{R}$ are orthogonal to each other, $\hat{\mathbf{Y}}'\mathbf{R} = 0$ and $\mathbf{R}'\hat{\mathbf{Y}} = 0$. This makes the above equation $\mathbf{Y}'\mathbf{Y} = \hat{\mathbf{Y}}'\hat{\mathbf{Y}} + \mathbf{R}'\mathbf{R}$. We can take the trace of this equation and the result is as follows:

$$tr(\mathbf{Y}'\mathbf{Y}) = tr(\hat{\mathbf{Y}}'\hat{\mathbf{Y}}) + tr(\mathbf{R}'\mathbf{R}) \tag{2.3}$$

This gives us the relationship between all of the sums of squares: $SST = SSTr + SSR$ where $SST = tr(\mathbf{Y}'\mathbf{Y})$, the treatment sum of squares $SSTr = tr(\hat{\mathbf{Y}}'\hat{\mathbf{Y}})$, and the residual sum of squares $SSR = (\mathbf{R}'\mathbf{R})$. In order to test the hypothesis mentioned above, the F test statistic is given by:

$$F = \frac{tr(\hat{\mathbf{Y}}'\hat{\mathbf{Y}})/(p-1)}{tr(\mathbf{R}'\mathbf{R})/(n-p)} \tag{2.4}$$

McArdle & Anderson (2001) propose a simple change to this theory that leads to a new non parametric MANOVA technique. This change takes advantage of the fact that tr(AB) = tr(BA) for any two matrices A and B where those matrix multiplications are compatible. So if we write equation 2.3 as $tr(\mathbf{YY}') = tr(\hat{\mathbf{Y}}\hat{\mathbf{Y}}') + tr(\mathbf{RR}')$, we can then solve for $\hat{\mathbf{Y}}\hat{\mathbf{Y}}' = \mathbf{HY}(\mathbf{HY})' = \mathbf{H}(\mathbf{YY}')\mathbf{H}' = \mathbf{H}(\mathbf{YY}')\mathbf{H}$. In the last step, $\mathbf{H}'$ becomes $\mathbf{H}$ since it is symmetric. Similarly, we get $\mathbf{RR}' = (\mathbf{1}-\mathbf{H})\mathbf{YY}'(\mathbf{1}-\mathbf{H})$. We can then replace $\mathbf{YY}'$ with Gower's centered matrix, a centered version of a squared distance matrix. This distance matrix could also be a semi-metric distance, which refers to a distance measure that might not necessarily satisfy the triangle inequality. To be specific, let $\mathbf{D} = (d_{ij})$ be some distance or semi-metric distance matrix, and $\mathbf{A} = (a_{ij}) = (-\frac{1}{2}d_{ij}^2)$. Then $\mathbf{G}$ is Gower's centered matrix given by:

$$\mathbf{G} = (\mathbf{I} - \frac{1}{n}\mathbf{1}\mathbf{1}')\mathbf{A}(\mathbf{I} - \frac{1}{n}\mathbf{1}\mathbf{1}') \tag{2.5}$$

This replaces $\mathbf{YY}'$ in the partitioning of the sum of squares which yields the analogy of Equation 2.3:

$$tr(\mathbf{G}) = tr(\mathbf{HGH}) + tr[(\mathbf{1}-\mathbf{H})\mathbf{G}(\mathbf{1}-\mathbf{H})] \tag{2.6}$$

Note that this is analogous to the sum of squares breakdown but it does not guarantee that the sums of squares are positive. The F test statistic is:

$$F = \frac{tr(\mathbf{HGH})/(p-1)}{tr[(\mathbf{1}-\mathbf{H})\mathbf{G}(\mathbf{1}-\mathbf{H})]/(n-p)} \tag{2.7}$$

The F statistic determined above is no longer the sum of independent chi-squared variables, nor is the assumption of normality necessarily valid due to the replacement of $\mathbf{YY}'$ by $\mathbf{G}$. Therefore tables cannot be used to find the p-value, rather a method of permutations is used. 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 (Anderson (2001)). The same rule applies to the two-way non-parametric MANOVA.

Anderson (2001) describes the one-way balanced non-parametric MANOVA as well as a two-way balanced MANOVA. From both of the papers mentioned, we were able to create a partial F-test in order to detect a row or column effect for data that are not necessarily balanced. Suppose we have two factors on the data: factor A which has a levels, and factor B which has b levels. In order to detect a treatment effect for factor A, let us consider a reduced model $\mathbf{Y} = \mathbf{X}_r\boldsymbol{\beta}_1 + \boldsymbol{\epsilon}$ where $\mathbf{X}_r$ is a design matrix that ignores factor A, and a full model $\mathbf{Y} = \mathbf{X}_f\boldsymbol{\beta} + \boldsymbol{\epsilon}$ where $\mathbf{X}_f$ is a design matrix involving both factors. Then the hypothesis to test for a treatment effect for factor A is $H_o : \boldsymbol{\beta}_1 = \mathbf{0}$. The F statistic for a partial F-test is given by:

$$F = \frac{(SSR_r - SSR_f)/(df_r - df_f)}{SSR_f/df_f} \tag{2.8}$$

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. When the difference is calculated between the two, it represents how many more parameters the full model has compared to the reduced model. From the derivations earlier in the section, we are able to input these sum of squares in terms of the trace of matrices as follows:

$$F = \frac{(tr(\mathbf{R}_r'\mathbf{R}_r) - tr(\mathbf{R}_f'\mathbf{R}_f))/(df_r - df_f)}{tr(\mathbf{R}_f'\mathbf{R}_f)/df_f} \tag{2.9}$$

We also know that since both models describe the data, we get that $tr(\mathbf{Y}'\mathbf{Y}) = tr(\hat{\mathbf{Y}}_r{}'\hat{\mathbf{Y}}_r) + tr(\mathbf{R}_r'\mathbf{R}_r)$ as well as $tr(\mathbf{Y}'\mathbf{Y}) = tr(\hat{\mathbf{Y}}_f{}'\hat{\mathbf{Y}}_f) + tr(\mathbf{R}_f'\mathbf{R}_f)$. Equating these and rearranging allows us to obtain $tr(\mathbf{R}_r'\mathbf{R}_r) - tr(\mathbf{R}_f'\mathbf{R}_f) = tr(\hat{\mathbf{Y}}_f{}'\hat{\mathbf{Y}}_f) - tr(\hat{\mathbf{Y}}_r{}'\hat{\mathbf{Y}}_r)$. Using this in equation 2.9 we obtain:

$$F = \frac{(tr(\hat{\mathbf{Y}}_f{}'\hat{\mathbf{Y}}_f) - tr(\hat{\mathbf{Y}}_r{}'\hat{\mathbf{Y}}_r))/(df_r - df_f)}{tr(\mathbf{R}_f'\mathbf{R}_f)/df_f} \tag{2.10}$$

Now, let $\mathbf{H}_f$ and $\mathbf{H}_r$ be the hat matrices defined above for the full and reduced models respectively, with $\mathbf{G}$ as defined above, then we can follow the steps in McArdle & Anderson (2001) to substitute into the equation 2.10 to obtain our final partial F statistic:

$$F = \frac{(tr(\mathbf{H}_f\mathbf{G}\mathbf{H}_f) - tr(\mathbf{H}_r\mathbf{G}\mathbf{H}_r))/(df_r - df_f)}{tr([\mathbf{I} - \mathbf{H}_f]\mathbf{G}[\mathbf{I} - \mathbf{H}_f])/df_f} \tag{2.11}$$

The same method for calculating the p-value described for the one-way MANOVA can be used to calculate the p-value for the two-way case.

Chapter 3

Simulation Study

3.1 Pseudo-Predators

Simulations based on realistic diets for seals are used to obtain the power and type I error rate of the MANOVA test described in Section 2.5. We refer to these simulated seals as pseudo-predators. These pseudo-predators have also been used in Iverson et al. (2004). In order to create these simulated seals, a prey database composed of the fatty acid signatures of samples of prey species is required. A specific diet is selected for the pseudo-predator, and fatty acid signatures of the species are sampled with replacement proportionate to the species included in the selected diet. A noise factor of 10% is included to represent small amounts of other prey that might be included in a real diet. The prey database used was collected on the Scotian Shelf off eastern Canada (Budge et al. (2002)). Also, as in Stewart & Field (2011), fat content is taken into account since prey with higher fat content will contribute more to the fatty acid signature of the predator. Calibration factors are also used to take into account that, for certain fatty acids, the quantity in the predator will always be higher or lower than the quantity in the prey. Only the fatty acids related to diet, or both diet and biosynthesis, are included in the signatures.

The diets selected for the simulations are based on realistic diets of seals on the east coast of Canada. The simulations for the one-way MANOVA used combinations of three of the six diets shown in Table 3.1. Diets chosen were selected based on the distances between the diets. In order to assess the performance of the one-way non-parametric MANOVA, we needed to simulate no effect (no difference in diets), a small effect (small distance between diets), and a large effect (large distance between diets). For no effect, all 3 diets were Diet 1, for a small effect we used Diet 1, Diet 2 and Diet 3, and for a large effect we used Diet 1, Diet 3 and Diet 5.

The simulations for the two-way MANOVA used combinations involving all 6 of the diets. Diets were selected similar to the one-way MANOVA only now considering

two factors. Therefore we wanted to select the diets to have a simulation with no effects, a small and large effect from factor A, a small and large effect from factor B, and one with an effect from both factors. The diets selected to accomplish this are shown in Table 3.3.

The distances between all of the diets are calculated using the chi-squared distance because of the presence of zeros. Because there is no sample of diets to calculate the appropriate γ, Stewart et al. ("In Review") suggests using $\gamma = \frac{1}{3}$ to compute the distance between diets as it corresponds to what was observed in practise. These distances are shown in Table 3.2. The effect size was calculated as the maximum chi-squared distance using $\gamma = \frac{1}{3}$ out of all possible pairwise distances between the diets used (shown in Table 3.2), as discussed above; alternative calculations are possible. The effect size when all three diets used were Diet 1 was found to be 0, the effect size when using Diet 1, Diet 2, and Diet 3 was found to be 0.8074 and the effect size when using Diet 1, Diet 3 and Diet 5 was found to be 1.973.

Species	Diet 1	Diet 2	Diet 3	Diet 4	Diet 5	Diet6
Northern Sandlance	35	38	42	45	50	53
Redfish	0	0	0	0	0	0
Capelin	0	0	0	0	0	0
Atlantic Cod	30	27	24	20	15	10
Silver Hake	15	12	10	10	5	5
American Plaice	0	0	0	0	0	0
Yellowtail Flounder	10	13	14	15	20	22
Longhorn Sculplin	0	0	0	0	0	0
Other	10	10	10	10	10	10
Total	100	100	100	100	100	100

Table 3.1: True diets of the pseudo-seals used in the simulations

	Diet 1	Diet 2	Diet 3	Diet 4	Diet 5	Diet 6
Diet 1	0					
Diet 2	0.4895	0				
Diet 3	0.8074	0.3385	0			
Diet 4	1.026	0.5853	0.2972	0		
Diet 5	1.973	1.500	1.171	0.9863	0	
Diet 6	2.356	1.901	1.583	1.337	0.5498	0

Table 3.2: Chi-squared distances between diets using $\gamma = 1/3$

3.2 Computing the simulations

The first step in simulations for the one-way MANOVA is to select 3 diets: π_1, π_2, and π_3. These selections are shown and described in Section 3.1. Samples of n_1, n_2, and n_3 pseudo-predators are then generated with diets π_1, π_2, and π_3 respectively. From these generated samples, a non-parametric MANOVA is computed and an F statistic and p-value are obtained. This process is repeated for M simulations. If all the diets are the same ($\pi_1 = \pi_2 = \pi_3$), then the probability of type I error is calculated as follows:

$$P[TypeIerror] = \frac{\#\{p.value \leq \alpha\}}{M} \tag{3.1}$$

where α is the targeted type I error rate or significance level. If at least one diet is different, then the power is calculated as follows:

$$Power = \beta(\pi_1, \pi_2, \pi_3) = 1 - \frac{\#\{p.value > \alpha\}}{M} \tag{3.2}$$

This process is repeated for different combinations of sample sizes, effect sizes, and distance measures. Note that in the simulated fatty acid signatures of the pseudo-predators, there are no zeros to modify in order to use Aitchison's distance. Two different sample sizes were used that are representative of nature: a small sample size ($n_1 = n_2 = n_3 = 15$) and a larger sample size ($n_1 = n_2 = n_3 = 45$).

For the two-way MANOVA simulations, the first step is to select six different diets, π_1, π_2, π_3, π_4, π_5, and π_6. Each diet is to be placed within a combination of two different factor groups A and B, where factor A has 2 levels and factor B has 3 levels. A sample of n pseudo-predators is then generated from each diet. These groups of generated pseudo-predators are then used to compute the two-way non-parametric MANOVA from which F statistics and p-values are obtained. This is repeated for M simulations. If the diets are the same, the type I error rate is calculated as in Equation 3.1 and if there is at least one diet that is different, the power is calculated as in Equation 3.2. This method is repeated for 6 different combinations of diets to assess the behaviour of the test when there are small and large effects due to factor A, factor B, both, and neither as discussed below. For all simulations, each group has a sample size of $n = 10$, making this a balanced two-way MANOVA, and only

Aitchison's distance is used because the chi-squared distance takes an unreasonable length of time to compute. This is due to the multitude of stress values and distances that require computing in order to choose γ. The different combinations of diets used for the simulations are shown below in Table 3.3 where the diet used for each combination of factor levels is shown for each simulation, given the diets described in Table 3.1.

Simulation	A\B	B Level 1	B Level 2	B Level 3
1	A Level 1	Diet 1	Diet 1	Diet 1
No effect	A Level 2	Diet 1	Diet 1	Diet 1
2	A Level 1	Diet 1	Diet 2	Diet 3
Small B effect	A Level 2	Diet 1	Diet 2	Diet 3
3	A Level 1	Diet 1	Diet 3	Diet 5
Large B effect	A Level 2	Diet 1	Diet 3	Diet 5
4	A Level 1	Diet 1	Diet 1	Diet 1
Small A effect	A Level 2	Diet 2	Diet 2	Diet 2
5	A Level 1	Diet 1	Diet 1	Diet 1
Large A effect	A Level 2	Diet 5	Diet 5	Diet 5
6	A Level 1	Diet 1	Diet 3	Diet 5
A and B effects	A Level 2	Diet 4	Diet 6	Diet 2

Table 3.3: Diet combinations used for the two-way simulations

Combinations were chosen strategically to see how well the two-way non-parametric MANOVA tests particular effects. The first simulation describes a situation with no effect since all the diets are the same. Simulations two and three represent the existence of a small and large effect respectively due to factor B. Simulations four and five represent the existence of a small and large effect respectively due to factor A. Finally, simulation six represents an effect due to both factors A and B. With these combinations, we are able to test the behaviour of the two-way non-parametric MANOVA for the plausible range of situations.

3.3 Simulation Results

To get the power and type I error results for the one-way MANOVA, 1000 simulations were completed for sample sizes of 15 and 45 using Aitchison's distance and the chi-squared distance explained in Chapter 2. Type I error rates for simulations using three equal diets (Diet 1) are shown in Table 3.4 as point estimates. The 99% confidence

intervals are computed for these point estimates to see whether the targeted type I error rates are within the intervals. These confidence intervals are based on a binomial confidence interval where the formula is as follows:

$$\hat{p} \pm Z_{1-\frac{1}{2}\alpha}\sqrt{\frac{1}{n}\hat{p}(1-\hat{p})} \tag{3.3}$$

Type I error rate α	n	Dist	Point Estimate	Lower Bound	Upper Bound
0.01	15	CHI	0.006	0	0.0123
		AIT	0.007	0.0002	0.0138
	45	CHI	0.013	0.0038	0.0222
		AIT	0.018	0.0072	0.0289
0.05	15	CHI	0.033	0.0184	0.0476
		AIT	0.043	0.0265	0.0595
	45	CHI	0.045	0.0281	0.0619
		AIT	0.050	0.0322	0.0678
0.1	15	CHI	0.082	0.0597	0.1043
		AIT	0.097	0.0729	0.1211
	45	CHI	0.094	0.0702	0.1178
		AIT	0.088	0.0649	0.1111

Table 3.4: Point estimates and 99% binomial confidence intervals for the type I error simulation results for 1000 simulations for 3 significance levels using sample sizes of 15 and 45 and with both the chi-squared distance measure and Aitchison's distance measure

It can be seen that when the small sample size is used ($n = 15$), the test behaves well, although not as well as with the larger sample size. This is important because often with fatty acid data, the sample sizes are small, so we need the tests to behave well both with larger and smaller sample sizes. In general, the test performs well for both sample sizes and for both distances as the targeted type I error rates (0.01, 0.05, 0.1) are within the 99% confidence intervals for almost all of the simulated type I error results shown in Table 3.4. The only interval that does not include the targeted type I error rate is for the sample size of 15 using the chi-squared distance with a targeted type I error rate of 0.05. This tells us that Aitchison's distance yields more accurate results at lower sample sizes.

The power results from the simulations were placed into two figures where each graph depicts power for the simulations based on one of the sample sizes used. The power results using $n = 15$ are shown in Figure 3.1 and the power results using $n = 45$

are shown in Figure 3.2.

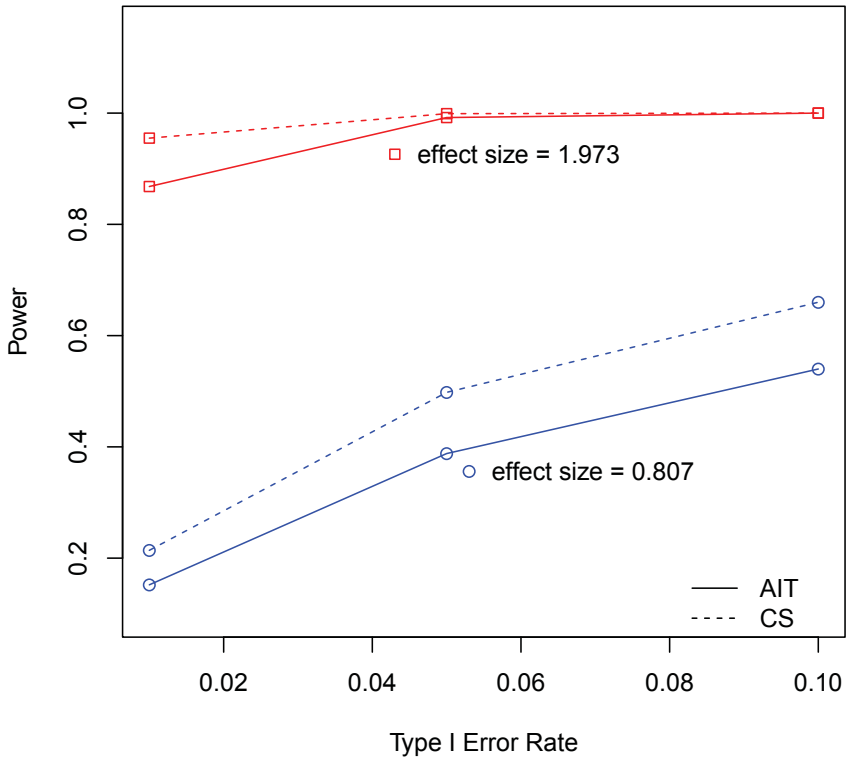

Figure 3.1: Power of simulations using sample size $n = 15$ and using both Aitchison's and chi-squared distances for various significance levels and effect sizes

From Figure 3.1, we can see that when the effect size is larger, the power increases, which is the desired effect. We can also see that the chi-squared distance gives higher power for both effect sizes. In Figure 3.2, the same conclusions can be reached. Comparing the two figures, we can see that the larger sample size yields a larger power which is expected.

The two-way simulations were computed according to the descriptions in Section 3.2. The type I error rates for factors A and B are shown for the simulations that had no difference in the diets for those factor groups. These values are given in Table 3.5 as point estimates, as well as the 99% binomial confidence interval for the type I error results.

We can see from Table 3.5 that the targeted type I error rates (0.01, 0.05, 0.1)

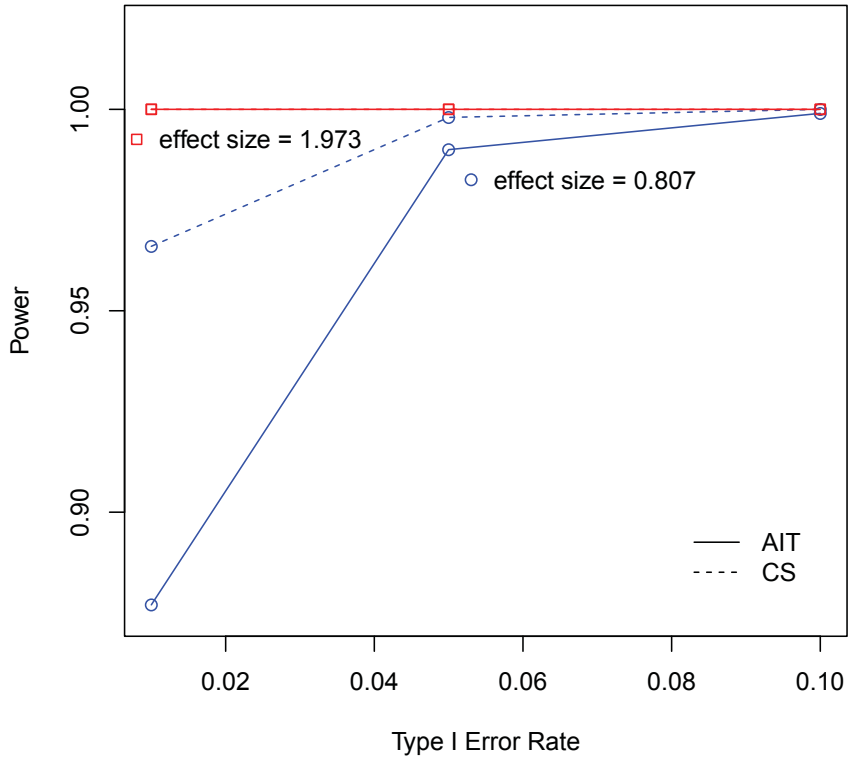

Figure 3.2: Power of simulations using sample size $n = 45$ and using both Aitchison's and chi-squared distances for various significance levels and effect sizes

are within the 99% confidence interval for most of the simulated type I error results. The exceptions include all of the simulation 4 results; factor B in simulation 1 for a significance level of 0.5 and 0.1; and simulation 2 for significance level 0.1. These results indicate that we are not rejecting the hypothesis of no A effect as often as expected. This could mean that it is harder to detect an effect due to factor A than it is to detect an effect due to factor B when there is only a small difference present. This may be due to the fact that there are three treatment groups in factor A, as opposed to only two in factor B, however it was expected that this would cause the opposite effect. This is something to be looked at in future research. Other than this, the two-way non-parametric MANOVA performs well when there is no effect in factor A and/or in factor B present.

The power results of the remaining simulations are calculated when there was an

Type I error rate α	Simulation	Factor	Point Estimate	Lower Bound	Upper Bound
0.01	1	A	0.015	0.0051	0.0249
		B	0.006	0	0.0123
	2	A	0.008	0.0007	0.0153
	3	A	0.012	0.0031	0.0209
	4	B	0.004	0	0.0091
	5	B	0.007	0.0002	0.0138
0.05	1	A	0.043	0.0265	0.0595
		B	0.031	0.0169	0.0451
	2	A	0.039	0.0232	0.0548
	3	A	0.052	0.0339	0.0701
	4	B	0.032	0.0177	0.0463
	5	B	0.037	0.0216	0.0524
0.1	1	A	0.087	0.0640	0.1100
		B	0.078	0.0562	0.0998
	2	A	0.075	0.0535	0.0965
	3	A	0.092	0.0685	0.1155
	4	B	0.061	0.0415	0.0805
	5	B	0.098	0.0738	0.1222

Table 3.5: Point estimates and 99% binomial confidence intervals for the type I error simulation results for 1000 simulations and 3 significance levels

effect due to one of the factors as represented by a difference in diets in one or both of the factors. Simulations 2 and 3 had different diets for factor B, simulations 4 and 5 had different diets for factor A, and simulation 6 had different diets for both factor A and factor B, indicating the presence of an effect due to these factors. The power results are shown in Figures 3.3 and 3.4 where the former includes the power results when there was an effect due to factor A and the latter includes the power results when there was an effect due to factor B.

From the figures, we can see again that factor A has a very low power of rejecting when the effect size is small compared to factor B, possibly be due to the number of treatments in each factor group. It is unclear why this causes lower power for factor A but it is something to be explored in future research. Also, it appears that simulation 6 for both factor A and B yields similar power results as the small effect size for each factor. We can see that the desired trait of increasing power as the significance level increases is achieved, as well as higher powers with a larger effect size. It should be noted, that because of the selection of the diets in simulation 6, it is more difficult to

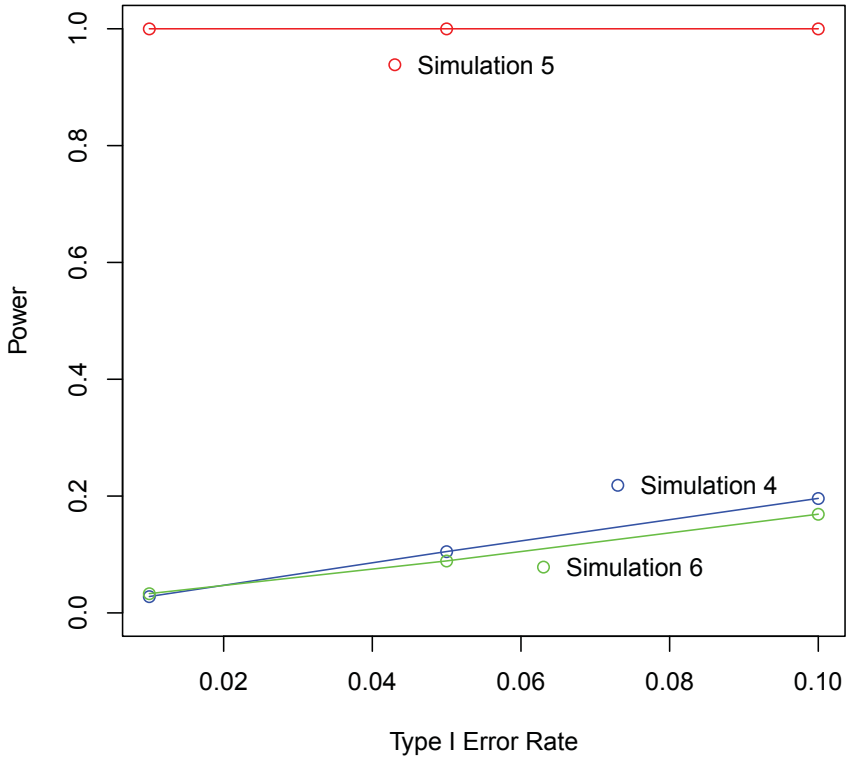

Figure 3.3: Power results for the two-way simulations for a factor A effect using various significance levels

observe an effect due to factor A because of the placement of Diet 5 and Diet 2. If these diets were interchanged, it would be easier to distinguish a difference between treatment 1 and treatment 2. In general, the one-way and the two-way MANOVAs perform well.

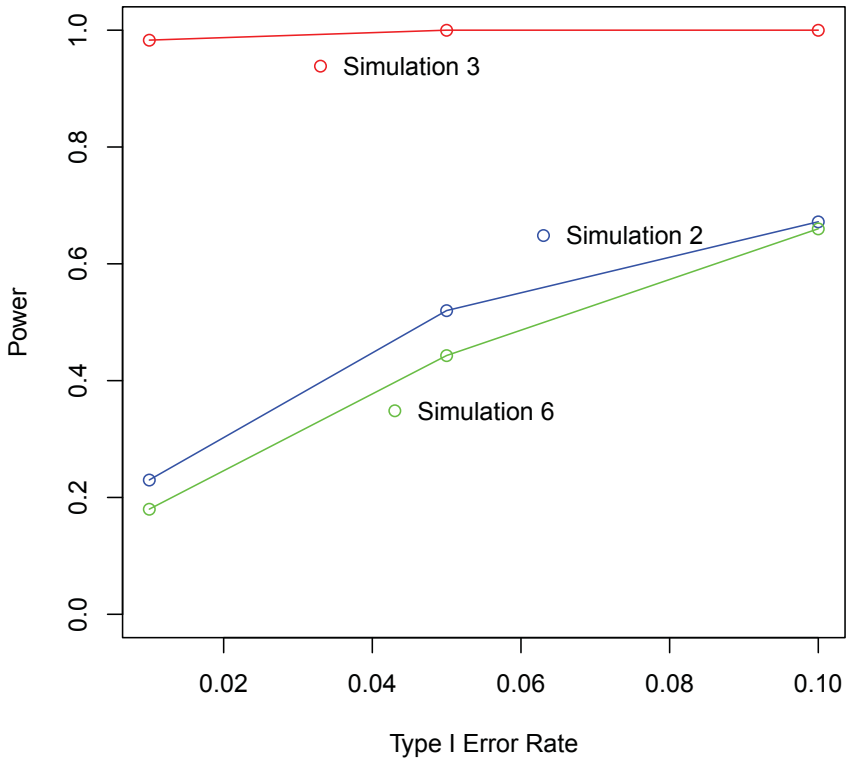

Figure 3.4: Power results for the two-way simulations for a factor B effect using various significance levels

Chapter 4

Results

Greenland sharks are a potentially abundant apex predator in the North Atlantic and Arctic Oceans that have recently generated ecological interest. As a species that has not been significantly studied to date, Greenland sharks are now being examined as they are potentially at risk due to increased fishing and climate change (MacNeil et al. (2012)). Climate change is warming the Arctic rapidly, causing a loss in seasonal and multi-year ice that is predicted to accelerate. The loss of seasonal ice is altering production in the dominant benthic food web and changing the availability and distribution of prey species. As an apex predator, Greenland sharks have the potential to be strongly affected by shifting prey distributions, changing their ecological role just as it is beginning to be understood.

Dietary information is important for understanding the structure of marine food webs and, as a large predator, the dietary compositions of Greenland sharks could help clarify the role they play in the Arctic. Because they are deep, cryptic predators, dietary fatty acid analysis has great potential for quantifying differences in dietary composition among Greenland shark populations that would otherwise be difficult to observe. To this end, fatty acid signatures of 112 sharks were recorded off Svalbard, Norway and Cumberland Sound, Canada during summer and winter from 2007 to 2009. Since a prey database is not available from either location, we cannot compute estimates for Greenland shark dietary compositions using QFASA, but we can test for a difference in fatty acid signatures between the two locations. If a significant difference exists, we can infer a difference in diets as we have no reason to believe that prey fatty acid signatures are different between locations. This will determine whether the sharks' role in the ecosystem near Svalbard is different from their role near Cumberland sound and could also explain the difference in physical characteristics between the two populations.

The data collected is depicted by the mean fatty acid signature at both locations

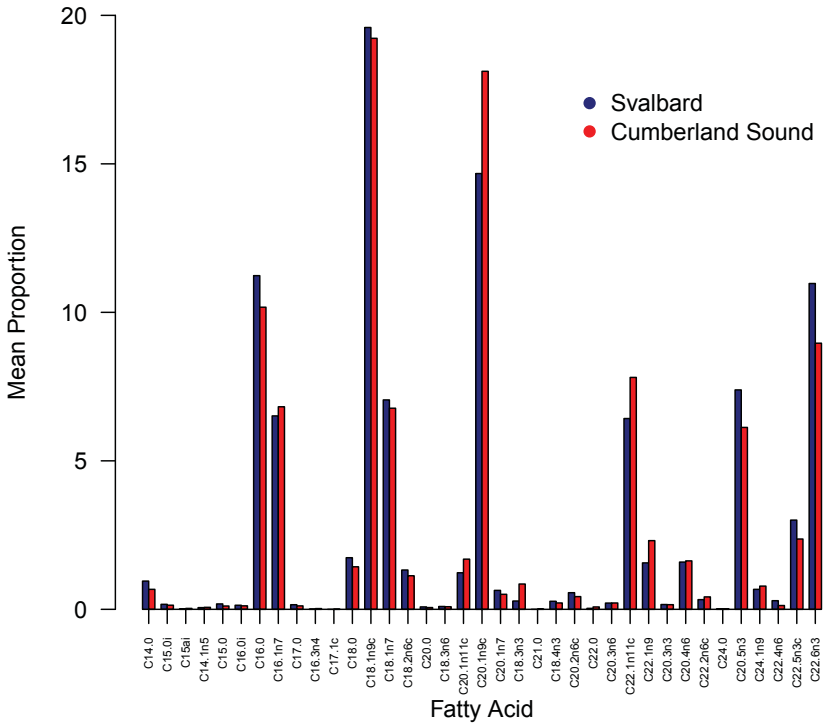

Figure 4.1: Sample mean fatty acid signatures of Greenland sharks at both Svalbard and Cumberland Sound

shown in Figure 4.1. We can see that the fatty acid signatures are similar in the fact that they have peaks in the same locations, but with small differences visible. In order to test whether this difference in shark fatty acid signatures at the two locations is significant, a non-parametric MANOVA described in Section 2.5 is used. If there is a significant difference, it gives strong evidence that there is a difference between diets at the two locations (Stewart et al. ("In Review")). The distance measure used for the MANOVA is the chi-squared distance described in Section 2.4.2. This distance measure was selected because when used with the non-parametric MANOVA, it tended to have a higher power than a MANOVA using Aitchison's distance. This was found using simulations described in Chapter 3. Using the method described in Section 2.4.2, a γ value of $\gamma = \frac{1}{11}$ was obtained. This γ was used for the chi-squared distance measure in order to compute the non-parametric MANOVA. The F-statistic

was found to be 8.712 giving a p-value of approximately 0. From this, we can conclude that there is strong evidence to say the fatty acid signatures of the Greenland sharks at Svalbard and Cumberland Sound differ and that the diets of the sharks at the two locations are likely different as well. When Aitchison's distance was used, the F-statistic was found to be 11.48 which also yields a p-value of approximately 0, leading us to the same conclusion.

Many factors could be causing the diets at the two locations to be different. There could simply be different distributions of prey with different fatty acid signatures, or the predators' ecological roles could be different. For instance, the Greenland sharks off of Svalbard are known to be up to 4 or 5 meters in length, whereas those off of Cumberland Sound tend to be approximately 1 or 2 meters shorter (Fisk et al. (2002), McMeans et al. (2013)). This size difference could allow them to feed on larger prey than the sharks off of Cumberland Sound. Also, as waters warm, marine species distributions are shifting northward and deepening (Brander (2010)). It is difficult to detect the magnitude of effects that these invading species have on the Arctic ecosystem to date, but they could be affecting prey availability and leading to different diets at the two locations.

We are also interested in testing whether there is a seasonal effect and/or a yearly effect on the diets of the sharks. As there are now three factors: season, year, and location, a three-way MANOVA is required. For this, the theory behind the two-way MANOVA was simply expanded to include a third factor. In the three-way MANOVA, we used Aitchison's distance because the calculations for the three-way MANOVA are already time consuming, and the chi-squared distance is much slower to calculate than Aitchison's distance. The F-statistic for location was found to be 9.110, which is close to that found when the other factors were not included. This once again yields a p-value of 0 telling us that there is strong evidence supporting the hypothesis that diets are different between locations. The F-statistics for testing for a seasonal effect and a yearly effect were found to be 10.349 and 14.826 respectively. These both yielded a p-value of approximately 0 allowing us to say that there is both a seasonal and a yearly effect on the diets of the sharks.

Seasonal effects have been previously explored in the Arctic. Changing water temperatures between seasons has been known to affect the distribution and abundance

of prey species (Hovde et al. (2002)). A large contribution to this seasonal change in diet could be due to the loss of seasonal sea ice during the summer, decreasing the quantity of ice connecting the multi-year ice to the open water. Algal mats form on the bottom of seasonal ice during the winter, that subsequently fall to the sea floor as a source of primary production for the benthic food web (MacNeil et al. (2010)). Therefore factors altering production in these regions seasonally likely alters the distribution of prey causing seasonal changes in Greenland shark diets.

Annual diet shifts have not been rigorously studied to date. However, the existence of such an effect is expected through climate change (MacNeil et al. (2012)). Although difficult to predict, it is thought that these changes will continue as the loss of sea ice increases every year. These changes may have important consequences for the ecosystem, including the distribution of prey species as warming waters cause a northward shift of temperate species distributions into the Arctic. However, such a strong difference from year to year was surprising, so we decided to look at the mean fatty acid signatures from Cumberland Sound for winter and summer during 2007, 2008 and 2009 in order to remove the other effects. The bar plot for winter 2008-2009 is shown in figure 4.2 and the bar plot for summer 2007-2009 is shown in figure 4.3.

From figure 4.2, we can see that certain fatty acids have very large relative differences including C22.1n11c, C20.1n11c, C22.5n3c, and C18.3n3. There are also some fatty acids that are present one year, but not the other, for instance C16.3n4, C20.0, and C24.0. Similarly, from figure 4.3, there are large relative differences between certain fatty acids including those mentioned above. Also, there are some fatty acids that disappear in 2009 after being present during 2007 and 2009 such as C22.0. These large relative differences between fatty acids from year to year could be potential contributors to the significant difference found in fatty acid signatures between the years.

Because we are interested in the difference in fatty acid signatures between locations while there is both a yearly and a seasonal effect present, we will look at data from only one season and year in order to eliminate the seasonal and yearly effects. Consider the data from both Cumberland Sound and Svalbard collected during summer 2008. The bar graph for the mean fatty acid signatures of the data collected during this period is shown in Figure 4.4.

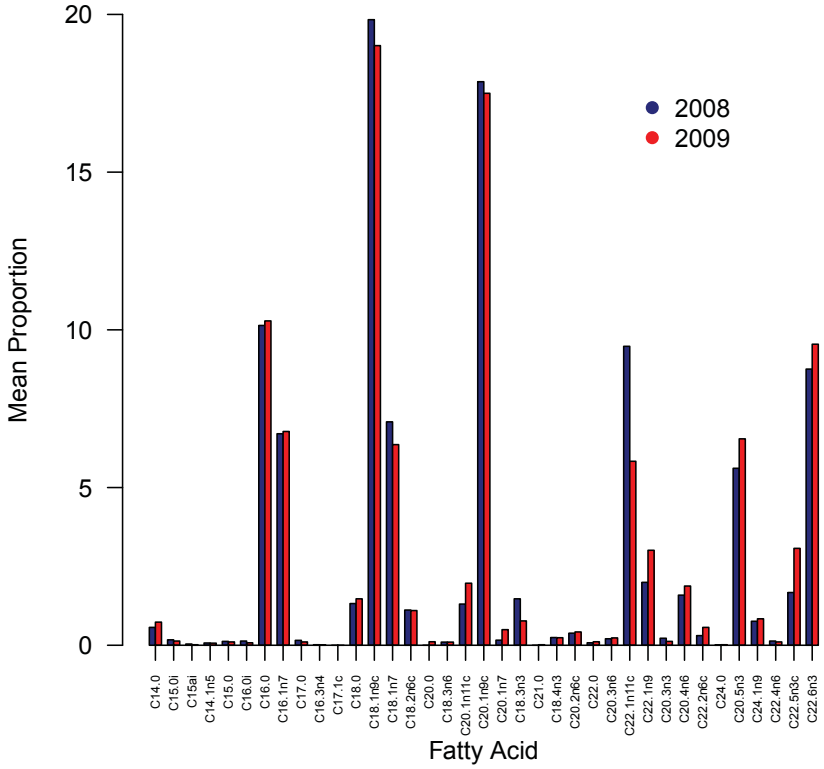

Figure 4.2: Sample mean fatty acid signatures of Greenland sharks in Cumberland Sound collected during winter 2008-2009

From the graph, we can see that there are several fatty acids that have large differences between the two locations such as C20.1n9c, C22.6n3, and C22.2n6c. But how large represents a significant difference between the fatty acids? For this, the relative difference between the two mean fatty acid signatures is calculated relative to the smaller proportion. So the relative difference is the smaller proportion subtracted from the larger proportion and divided by the smaller proportion. These values are then used to graph a box plot shown in Figure 4.5.

This box plot is graphed with outliers, which were fatty acids C15ai, C16.3n4, C18.3n3, C22.2n6c and C24.0 with relative differences of 2.41, 16.07, 1.69, 11.22 and infinity respectively. All of these outlying relative differences had higher values in Cumberland Sound than in Svalbard. The upper fence was found to be 1.22 so any relative difference above 1.22 was found to be an outlier and therefore those fatty

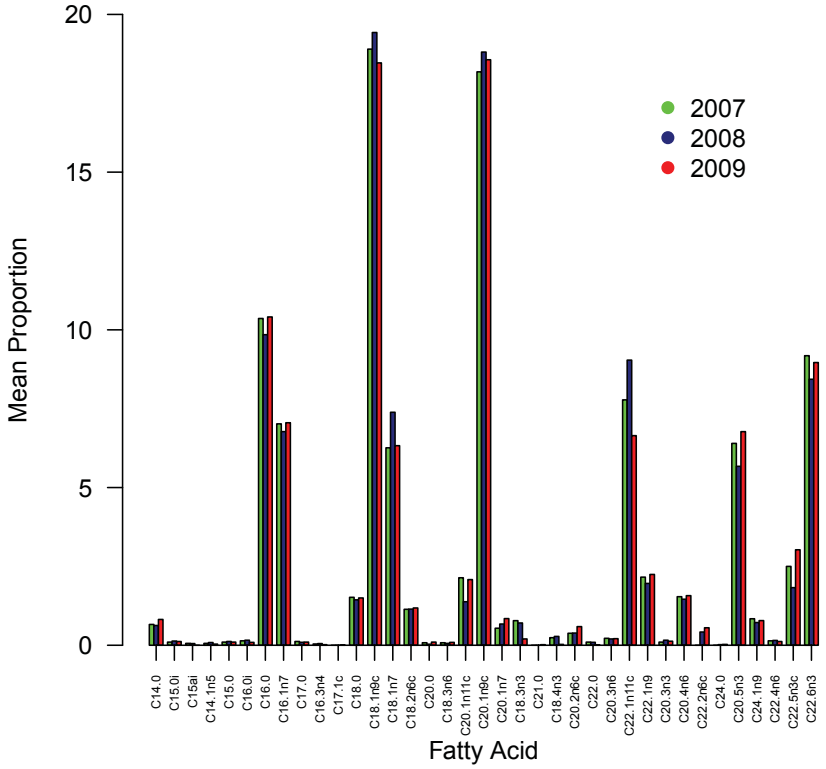

Figure 4.3: Sample mean fatty acid signatures of Greenland sharks in Cumberland Sound collected during summer 2007-2009

acids are different. This tells us that the proportions of these fatty acids are different and are potential contributors to the conclusion that the fatty acid signatures between the two locations differ. Although it hasn't been completed, it may be valuable to remove these outlying fatty acids, reclose the compositions, and complete the analysis again. If we find that there is no longer a difference in fatty acid signatures, we could conclude that these fatty acids are causing the difference.

Potential sources for these differences are difficult to predict, however existing knowledge about these two ecosystems as well as fatty acid content in prey can help to explain some of the variation. Greenland sharks near Svalbard have been found to feed on mammals such as grey seals and minke whales that could be an important prey in their diet (Leclerc et al. (2011)). Whale blubber and offal is discarded into the ocean near Svalbard from whaling, with the sharks rushing to the surface to

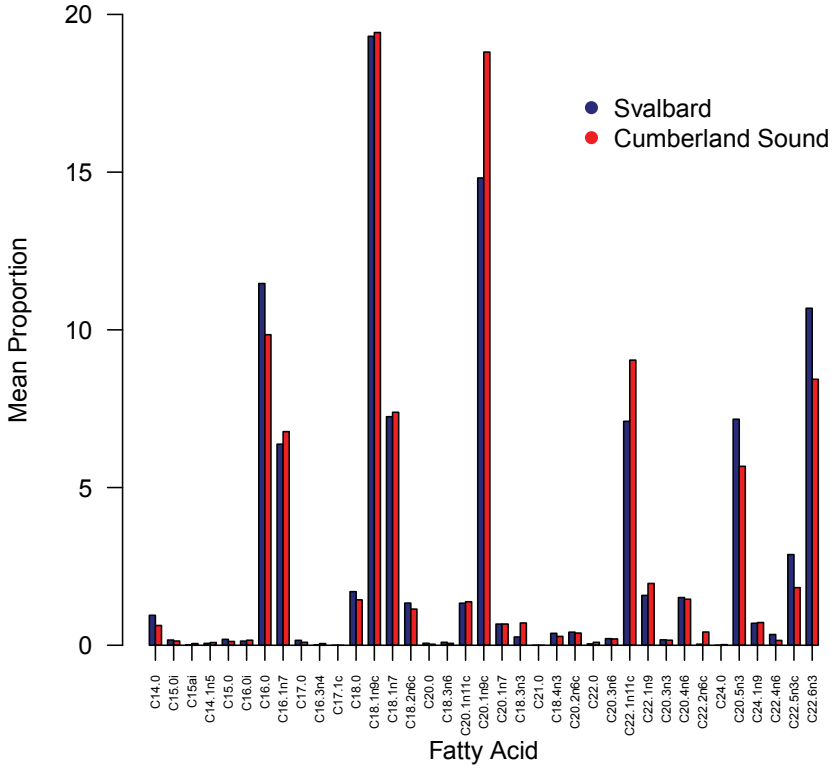

Figure 4.4: Sample mean fatty acid signatures of Greenland sharks at both Svalbard and Cumberland Sound collected during summer 2008

consume them, providing high energy food. Therefore consumption of this mammal based, high energy diet (namely ringed seals, Atlantic cod, and whale blubber) by Greenland sharks off Svalbard compared to a diet rich in Greenland Halibut for the sharks in Cumberland Sound is likely driving the differences in fatty acid signatures between locations (McMeans et al. (2013)).

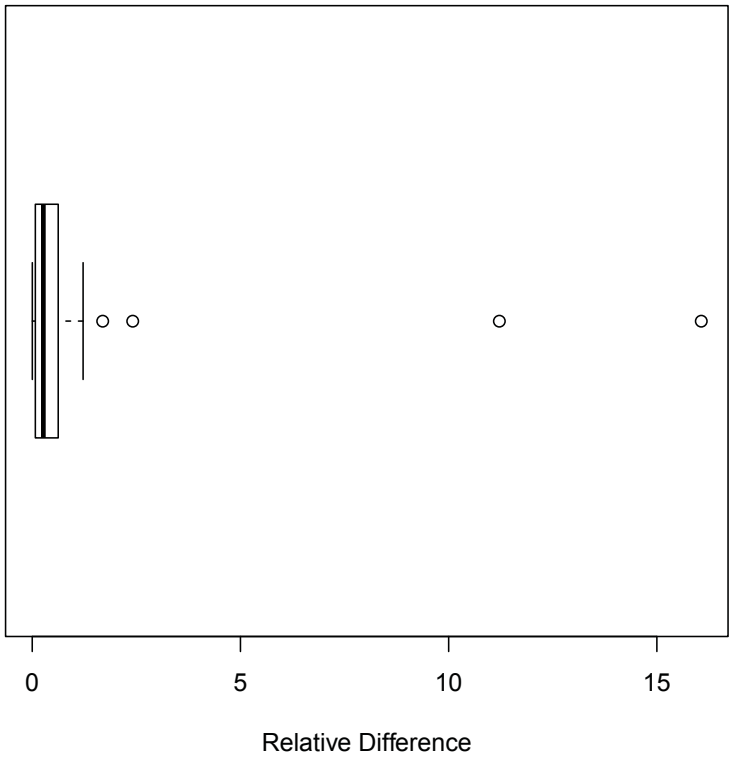

Figure 4.5: Box plot of the relative differences between the mean fatty acid signatures of Greenland Sharks at Svalbard and Cumberland Sound collected during summer 2008

Chapter 5

Conclusions

5.1 Summary

Our research was motivated by the desire to test whether there is a significant difference in the diets of Greenland sharks off Svalbard versus Cumberland Sound. Fatty acid signatures of the sharks are collected at these locations to study their diets. If a significant difference in the fatty acid signatures is found, it suggests a difference in diet (Stewart et al. ("In Review")). However because the number of dietary fatty acids is larger than the sample size, standard MANOVA processes cannot be used to test for this difference. A non-parametric MANOVA proposed by McArdle & Anderson (2001) is used.

The non-parametric MANOVA was based on matrix multiplications from a standard MANOVA and the transpositions of matrices. The sum of squares in a standard MANOVA is the trace of the matrix of interest (for example the residual matrix), multiplied by its transpose. Since it is always true that $tr(AB) = tr(BA)$ when the matrix multiplications between A and B are applicable, we can switch the order of the matrices within the trace function. This one change leads to a series of simplifications (see Section 2.5) that allow us to substitute Gower's centered distance matrix for the data matrix Y multiplied by its transpose Y' (YY'). This substitution allows us to use any distance measure that is appropriate for the type of data at hand. Also, because the formula for the non-parametric MANOVA statistic no longer uses the predicted values, nor the data matrix, it eliminates the need for the assumptions that sample size is larger than the number of parameters and that data are approximately normal. Because these assumptions are no longer required, the non-parametric MANOVA can be used on fatty acid data, specifically that of the Greenland sharks described above.

The p-value for this non-parametric MANOVA is based on permutations, since we can no longer assume that the statistic follows an F distribution. Since it is often unrealistic to calculate all possible permutations, a random sample of at least

1000 permutations can be used for a test with a level of significance of 0.05. These permutations are computed, the F-statistic is found for each permutation, and the p-value is given by the number of permuted F-statistics that are equal to or larger than the original F statistic divided by the number of permutations.

There is no restriction on which distance or semi-metric distance is used to calculate the F-statistic, but if a Euclidean distance measure is used and the assumption of normality of the data is valid, then this non-parametric method will yield the same results as the standard MANOVA technique. In this thesis, both Aitchison's distance and the chi-squared distance were explored. Aitchison's distance (defined in Equation 2.1) is commonly used on compositional data, but it does not work when there are zero elements in the data. In this case, a replacement method, like that proposed by Martín-Fernández & Thió-Henestrosa (2006), is required to replace the zeros with positive elements in order to use Aitchison's distance. An alternative to replacing the zeros is to use a distance capable of handling zeros. One such distance is the chi-squared distance (Equation 2.2) which is originally proposed in Greenacre (2011) and expanded on in Stewart et al. ("In Review").

These distances are used in both the one-way non-parametric MANOVA as well as the two-way. A partial F-test is proposed in order to complete a non-parametric two-way MANOVA based on the same sum of squares formulas as in McArdle & Anderson (2001). It combines the standard formula for a partial F-test (Equation 2.8) and the same variation in the sum of squares as described earlier, to obtain the non-parametric partial F-test (Equation 2.11). To examine the behaviour of these tests, simulations were computed.

5.1.1 Simulations

First, simulations were computed for the one-way non-parametric MANOVA. To compute the simulations, simulated predators are required. These simulated predators are called pseudo-predators and use a database of prey fatty acid signatures to simulate an individual pseudo-predator's fatty acid signature. A random sample of fatty acid signatures is selected with replacement from the prey database proportionate to the percentage of prey in the desired diet of the pseudo-predator. Noise can also be taken into account which represents other prey that real seals could be eating not listed in

the diet. The pseudo-predator is then the summary fatty acid signature (typically the mean) of this sample.

For one-way non-parametric MANOVA simulations, three diets are selected and a sample of pseudo-predators is generated from each. Diets were selected so that one simulation had all of the same diets, one had slightly different diets, and one had extremely different diets. This was to vary effect size when analyzing the power and type I error rate of the test. These diets are described in Table 3.1. Simulations were performed both with large and small sample sizes, using both Aitchison's distance measure and the chi-squared distance measure. Simulation results demonstrated that these tests perform well both with large and small sample sizes. The test also gave the desired effect that power increases with effect size. We also discovered that the chi-squared distance tends to yield larger power than Aitchison's distance. Although it gives more power to the test, when the two-way simulations were computed, we used Aitchison's distance because the chi-squared distance takes a very long time to compute due to the intensive calculations to choose γ. Power results using both the distances were close enough that we can reasonably use Aitchison's distance for the more computationally intense two-way MANOVA.

Two-way non-parametric MANOVA simulations were performed in a similar fashion. Six diets were selected and placed into different combinations of row factors and column factors. They were chosen in such a way so as to examine no effect, a small effect and a large effect due to factor A, due to factor B, and when there is an effect due to both factors A and B. The combinations of diets are described in Table 3.3. Simulations were done as a balanced two-way MANOVA using a sample size of 10 in each combination of factor groups. The two-way simulation results showed that the test performs well with small sample sizes, although it was more difficult to detect an effect for a treatment with two groups. It is unclear why this was happening and should be looked at in future research. Other than this, the test performs well when there are no effects as the type I error results were close to the targeted type I error rates, as well as when effects exist as the test showed strong power when the effect sizes were larger.

5.1.2 Ecological Data

The fatty acid data for Greenland sharks off Svalbard and Cumberland Sound was the main motivator for this thesis. The data set was collected with the intention of quantifying the difference in diet between the sharks at the two locations. It is hypothesized that this is due to differences in prey availability off of Svalbard compared to Cumberland Sound. We are interested in testing this, however because of the restrictions that compositional and fatty acid data present, a standard MANOVA technique cannot be used. These restrictions include that each individual fatty acid signature sums to 1, contains non-negative entries, and that often the number of parameters is larger than the sample size. In order to deal with these restrictions, a non-parametric MANOVA is used.

When location was the only factor of interest, the one-way MANOVA test using chi-squared distance told us we have strong evidence to support our hypothesis that there is a difference in diets between the sharks at the two locations. If we used Aitchison's distance, the same conclusion was reached. This supports the suggestion that the difference in physical characteristics (for example, size of the sharks) is related to a difference in diets.

When we wanted to test for individual effects while considering all three factors (season, year, and location), a three-way MANOVA was required. This is carried out the same way as the two-way MANOVA only with one extra factor group. In other words, if we wanted to test for a seasonal effect, our reduced design matrix X_r would simply be the full design matrix without the columns pertaining to the seasonal factors. Similarly, the reduced design matrices for testing for year and then for location can be obtained. The partial F-test is then calculated separately for each factor with the reduced design matrices, and F-statistics and p-values are obtained.

F-statistics for seasonal effect, yearly effect and location effect were found to all be significantly large, yielding p-values of approximately zero. This tells us that there is a yearly effect, a seasonal effect, and a location effect on the fatty acid signatures of Greenland sharks. We have no reason to believe that this is attributed to any other difference than that of diets, therefore we can say that we have strong evidence supporting our hypothesis that the varying characteristics in the Greenland sharks at the two locations is related to differing diets. The seasonal effect is most likely due

to seasonal migrations of sharks and prey. A yearly effect isn't a topic that has been studied thus far, so it could be something of interest for future research.

5.2 Future Research

Conclusions have been made on the Greenland sharks' diets based solely on the fatty acid signatures of these sharks. In Stewart et al. ("In Review"), it is suggested that inferences on the diets of the sharks based solely on comparing fatty acid signatures can be made if the data is collected in the same season and same region. Here, the data has been collected over several seasons and at two different locations. Because we have no reason to believe that the fatty acid signatures of the prey are different at the two locations, we assume that a difference in fatty acid signatures is attributed to a difference in diet. It is possible however that the sharks do have the same diets, and the significant difference is caused by different fatty acid signatures of prey. In order to examine this further, a database of prey fatty acid signatures could be sampled at each location, and QFASA could be performed. Then, a non-parametric MANOVA could be performed on the diets themselves to test whether the change is due to different diets. The same strategy could be performed in order to test for a seasonal and a yearly effect.

Using diet estimates to test for a difference in dietary composition has several drawbacks. First, collecting a prey database is often a huge undertaking as it requires collecting many samples of prey and determining the fatty acid signatures for each – a lengthy process. Also, in order to obtain diet estimates, QFASA is required, which has some methods that could be developed further. For instance, in order to obtain more accurate results, calibration factors are used since for certain fatty acids, the amount in the predator may always be higher or lower than the amount found in the prey. These factors are estimated through a long-term experiment during which predators are fed a constant diet (Stewart & Field (2011)). The coefficients can then be estimated by the differences between fatty acid signatures in the predator compared with those of the prey. Further development on the estimation of calibration factors could help increase the accuracy of diet estimates obtained from QFASA which would also increase the effectiveness of testing for differences among diets. Similarly, fat content estimates are used in QFASA. These fat contents are based on the average

fat content of a sample of prey species, however there is large variation in fat content among some species. For these reasons, the development of a new method that included variation in fat content and calibration factors within a species is desired to obtain more accurate diet estimates.

As mentioned before, there has not been much research to date on a yearly effect on Greenland shark diets. This could be another consequence of warming waters due to climate change, which is predicted to have a particularly large effect on Arctic ecosystems. Gathering this data over many more years, as well as samples of prey from year to year, could help discover changes in prey distributions, fatty acid signatures, and dietary compositions of predators which could be attributed to climate change.

Appendix A

Chi-Squared Distance

```
choose.alpha.sub2 <- function(Y.1,Y.2) {

# NOTE, HERE "ALPHA" IS REFERRING TO 1/GAMMA
# FOR VARIOUS VALUES OF ALPHA, COMPUTES THE AVERAGE STRESS OVER ALL
# SUBCOMPOSITIONS OF SIZE 2

stress.avg <- numeric(50)

for(i in 1:50){

alpha <- i

# TRANSFORMING THE COMPOSITIONS

      Y.1.t <- Y.1^(1/alpha)
      Y.1.t <- Y.1.t/apply(Y.1.t,1,sum)

      Y.2.t <-  Y.2^(1/alpha)
      Y.2.t <- Y.2.t/apply(Y.2.t,1,sum)

      d.mat <- alpha * create.d.mat(Y.1.t,Y.2.t)*sqrt(ncol(Y.1))

      stress <- 0

      for (j in 1:(ncol(Y.1)-1)){
for (k in (j+1):ncol(Y.1)){
```

```
# COMPUTING TWO PART SUBCOMPOSITIONS
                S.1 <- Y.1[,c(j,k)]

                S.2 <- Y.2[,c(j,k)]

# CHECK THAT BOTH COMPONENTS ARE NOT ZERO

                both.zero.S1.vec <- (S.1[,1]==0) & (S.1[,2]==0)

                S.1[both.zero.S1.vec,] <- 1e-6

                S.1 <- S.1/apply(S.1,1,sum)

                both.zero.S2.vec <- (S.2[,1]==0) & (S.2[,2]==0)

                S.2[both.zero.S2.vec,] <- 1e-6

                S.2 <- S.2/apply(S.2,1,sum)

                S.1.t <- S.1^(1/alpha)
                S.1.t <- S.1.t/apply(S.1.t,1,sum)

                S.2.t <- S.2^(1/alpha)
                S.2.t <- S.2.t/apply(S.2.t,1,sum)

                s.mat <- alpha *create.d.mat(S.1.t,S.2.t) * sqrt(ncol(S.1))

                stress <- stress + stress.meas(d.mat,s.mat)
```

```
}
}

           stress.avg[i] <- stress/ncol(combn(ncol(Y.1),2))
}

min.stress <- min(stress.avg)
loc <- stress.avg == min.stress
fin.alpha <- seq(1,50,1)[loc]
return(fin.alpha)
}

choose.alpha.sub3 <- function(Y.1,Y.2,Y.3) {

# NOTE THAT ALPHA HERE IS 1\GAMMA
# THERE ARE NO ZEROS IN THE DATA
# FOR VARIOUS VALUES OF ALPHA, COMPUTES THE AVERAGE STRESS OVER ALL
# SUBCOMPOSITIONS OF SIZE 2 WHEN THERE ARE THREE COMPOSITIONS

stress.avg <- 10000
alpha <- 0

while (stress.avg > 1e-6) {

alpha <- alpha + 1

      # TRANSFORMING THE DATA

      Y.1.t <- Y.1^(1/alpha)
      Y.1.t <- Y.1.t/apply(Y.1.t,1,sum)

      Y.2.t <-  Y.2^(1/alpha)
```

```
      Y.2.t <- Y.2.t/apply(Y.2.t,1,sum)

      Y.3.t <-  Y.3^(1/alpha)
      Y.3.t <- Y.3.t/apply(Y.3.t,1,sum)

      d.mat <- alpha * create.d.mat2(Y.1.t,Y.2.t,Y.3.t)*sqrt(ncol(Y.1))

      stress <- 0

    for (j in 1:(ncol(Y.1)-1)){
        for (k in (j+1):ncol(Y.1)){

# COMPUTING TWO PART SUBCOMPOSITIONS

                  S.1 <- Y.1[,c(j,k)]

                  S.2 <- Y.2[,c(j,k)]

      S.3 <- Y.3[,c(j,k)]

                  S.1.t <- S.1^(1/alpha)
                  S.1.t <- S.1.t/apply(S.1.t,1,sum)

                  S.2.t <- S.2^(1/alpha)
                  S.2.t <- S.2.t/apply(S.2.t,1,sum)

      S.3.t <- S.3^(1/alpha)
                  S.3.t <- S.3.t/apply(S.3.t,1,sum)

                  s.mat <- alpha *create.d.mat2(S.1.t,S.2.t,S.3.t) * sqrt(ncol(S.1

                  stress <- stress + stress.meas(d.mat,s.mat)
```

```
}
}
stress.avg <- stress/ncol(combn(ncol(Y.1),2))
}
return(alpha)
}

create.d.mat <- function(Y.1,Y.2) {

# TWO COMPOSITIONS
# WANT TO CREATE A MATRIX OF DISTANCES (NOTE THAT d11!=0 AND d12!=d21 SINCE NOT TH
# 1 AND 2 ARE NOT THE SAME SEALS.

ns1 <- nrow(Y.1)
ns2 <- nrow(Y.2)
nFA <- ncol(Y.1)

ind.vec <- as.vector(unlist(tapply(seq(1,ns1,1),seq(1,ns1,1),rep,ns2)))

Y.1.rep <- Y.1[ind.vec, ]

Y.2.rep <- rep(t(Y.2),ns1)
Y.2.rep <- matrix(Y.2.rep,ncol=ncol(Y.2),byrow=T)

Y.1.split <- split(Y.1.rep,seq(1,nrow(Y.1.rep),1))
Y.2.split <- split(Y.2.rep,seq(1,nrow(Y.2.rep),1))

d.mat <- matrix(mapply(chisq.CA,Y.1.split,Y.2.split),byrow=T,ns1,ns2)
return(d.mat)
}

create.d.mat2 <- function(Y.1,Y.2,Y.3) {
```

```
# THREE COMPOSITIONS
# WANT TO CREATE A MATRIX OF DISTANCES (NOTE THAT d11!=0 AND d12!=d21 SINCE NOT TH
# 1 AND 2 ARE NOT THE SAME SEALS.

ns1 <- nrow(Y.1)
ns2 <- nrow(Y.2)
ns3 <- nrow(Y.3)
nFA <- ncol(Y.1)

ind.vec <- as.vector(unlist(tapply(seq(1,ns1,1),seq(1,ns1,1),rep,ns2)))
ind.vec2 <- as.vector(unlist(tapply(seq(1,ns1,1),seq(1,ns1,1),rep,ns3)))
ind.vec3 <- as.vector(unlist(tapply(seq(1,ns2,1),seq(1,ns2,1),rep,ns3)))

Y.1.rep2 <- Y.1[ind.vec, ]

Y.2.rep1 <- rep(t(Y.2),ns1)
Y.2.rep1 <- matrix(Y.2.rep1,ncol=ncol(Y.2),byrow=T)

Y.1.rep3 <- Y.1[ind.vec2, ]

Y.3.rep1 <- rep(t(Y.3),ns1)
Y.3.rep1 <- matrix(Y.3.rep1,ncol=ncol(Y.3),byrow=T)

Y.2.rep3 <- Y.2[ind.vec3, ]

Y.3.rep2 <- rep(t(Y.3),ns2)
Y.3.rep2 <- matrix(Y.3.rep2,ncol=ncol(Y.3),byrow=T)


Y.1.split2 <- split(Y.1.rep2,seq(1,nrow(Y.1.rep2),1))
Y.2.split1 <- split(Y.2.rep1,seq(1,nrow(Y.2.rep1),1))
```

```
Y.3.split1 <- split(Y.3.rep1,seq(1,nrow(Y.3.rep1),1))
Y.1.split3 <- split(Y.1.rep3,seq(1,nrow(Y.1.rep3),1))
Y.2.split3 <- split(Y.2.rep3,seq(1,nrow(Y.2.rep3),1))
Y.3.split2 <- split(Y.3.rep2,seq(1,nrow(Y.3.rep2),1))

d.mat1 <- matrix(mapply(chisq.CA,Y.1.split2,Y.2.split1),byrow=T,ns1,ns2)
d.mat2 <- matrix(mapply(chisq.CA,Y.1.split3,Y.3.split1),byrow=T,ns1,ns3)
d.mat3 <- matrix(mapply(chisq.CA,Y.2.split3,Y.3.split2),byrow=T,ns2,ns3)

d.vec <- as.matrix(c(as.vector(d.mat1),as.vector(d.mat2),as.vector(d.mat3)),nrow=1
return(d.vec)
}

chisq.CA <- function(x1,x2) {

# COMPUTES THE CHISQUARE DISTANCE SIMILAR TO THE ONE DISCUSSED
# IN (PAWLOWSKY-GLAHN AND BUCCIANTI, 2011)

d.sq <- (x1-x2)^2
c.vec <- x1+x2

if (any(d.sq!=0)) {

d.sq[d.sq!=0] <- d.sq[d.sq!=0]/c.vec[d.sq!=0]

}

d.sq <- 4*sum(d.sq)


d <- sqrt(d.sq/2)
```

```
return(d)
}


stress.meas <- function(d.mat,s.mat) {

# d.mat AND s.mat ARE CREATED FROM create.d.mat

# THIS FUNCTION COMPUTES THE STRESS BY COMPARING
# THE NUMBER OF DISTANCE MEASURES IN SUBCOMPOSITIONS (s.mat)
# THAT ARE LESS THAN OR EQUAL DISTANCE MEASURES in d.mat.

## GREENACRE (2011)
#stress <- sqrt( sum((d.mat - s.mat)^2)/sum(d.mat^2) )

ns1 <- nrow(d.mat)
ns2 <- ncol(d.mat)
stress <- 1- (sum(d.mat>=s.mat)/(ns1*ns2))

return(stress)
}

chisqr <- function(Y.1, Y.2, alpha){
#Y.1 and Y.2 are vectors

Y.1trans <- (Y.1^(1/alpha))/sum(Y.1^(1/alpha))
Y.2trans <- (Y.2^(1/alpha))/sum(Y.2^(1/alpha))

chisqr <- alpha*chisq.CA(Y.1trans, Y.2trans)*sqrt(length(Y.1))
return(chisqr)
}
```

Appendix B

Non-Parametric MANOVA

```
# load robCompositions
library(robCompositions)

npFtest2 <- function(seals, m, n1, n2, n3){

    # ONE-WAY MANOVA FOR 3 GROUPS USING AITCHISON'S DISTANCE
# seals = A MATRIX WHERE ALL THE GROUPS ARE rbind TOGETHER
# m = NUMBER OF GROUPS (m=3)
# n1, n2, n3 = NUMBER OF SEALS IN EACH GROUP RESPECTIVELY

n <- n1 + n2 + n3 # total number of observations

# creating the design matrix
X <- matrix(0, n, m)
X[,1] <- 1
X[1:n1,2] <- 1
X[(n1+1):(n1+n2),3] <- 1

d <- matrix(0, nrow=n, ncol=n)
# USING AITCHISON'S DISTANCE MEASURE aDist, CALCULATING A DISTANCE MATRIX

for(i in 1:n){
for(j in 1:n){
d[i,j] <- aDist(seals[i,],seals[j,])
}
}
```

```
dsqr <- d^2
A <- (-1/2)*dsqr

H <- X%*%solve(t(X)%*%X)%*%t(X)

ones <- matrix(1,n,1)
I <- diag(n)
g <- (I - (1/n)*ones%*%t(ones))
G <- g%*%A%*%g

F <- (sum(diag(H%*%G%*%H))/(m-1))/(sum(diag((I-H)%*%G%*%(I-H)))/(n-m))
return(F)
}


p.value2 <- function(seals, m, n1, n2, n3, nperm=1000){

    # FOR THREE COMPOSITIONS USING AITCHISONS DISTANCE
# seals = EACH ROW IS A SEAL, THE GROUPS OF SEALS ARE rbind TOGETHER
# m = THE NUMBER OF GROUPS (m=3)
# n1, n2, n3 = THE NUMBER OF SEALS IN EACH GROUP RESPECTIVELY
# nperm = THE NUMBER OF PERMUTATIONS TO COMPLETE FOR THE P-VALUE

Forig <- npFtest2(seals, m, n1, n2, n3)

# CALCULATING nperm RANDOM PERMUTATIONS AS DESCRIBED IN MARTI ANDERSON'S PAPER
# 1000 FOR 0.05 SIGNIFICANCE, 5000 FOR 0.01 SIGNIFICANCE
```

```
Fperm <- numeric(nperm)

for(i in 1:nperm){

ind <- c(1:nrow(seals))
perm <- sample(ind)
seals.perm <- seals[perm,]

Fperm[i] <- npFtest2(seals.perm, m, n1, n2, n3)
}

# HOW MANY PERMUTATIONS WERE GREATER THAN THE ORIGINAL Forig
p.value <- sum(Fperm > Forig)/nperm
return(p.value)
}

npFtest4 <- function(seals, n1, n2, alpha){

    # FOR TWO COMPOSITIONS USING CHI-SQUARED DISTANCE
# seals = A MATRIX WHERE ALL THE GROUPS ARE rbind TOGETHER
# n1, n2 = NUMBER OF SEALS IN EACH GROUP RESPECTIVELY
n <- n1 + n2

# CREATING THE DESIGN MATRIX
m <- 2
X <- matrix(0, n, 2)
X[,1] <- 1
X[1:n1,2] <- 1

d <- matrix(0, nrow=n, ncol=n)
# USING THE CHI-SQUARED DISTANCE MEASURE, CALCULATING A DISTANCE MATRIX
```

```
#getting the alpha for the chi-squared distance
#alpha <- choose.alpha.sub2(seals[1:n1,], seals[(n1+1):n,])

for(i in 1:n){
for(j in 1:n){
d[i,j] <- chisqr(seals[i,],seals[j,], alpha)
}
}

dsqr <- d^2
A <- (-1/2)*dsqr

H <- X%*%solve(t(X)%*%X)%*%t(X)

ones <- matrix(1,n,1)
I <- diag(n)
g <- (I - (1/n)*ones%*%t(ones))
G <- g%*%A%*%g


F <- (sum(diag(H%*%G%*%H))/(m-1))/(sum(diag((I-H)%*%G%*%(I-H)))/(n-m))
return(F)
}


p.value4 <- function(seals, n1, n2, alpha, nperm=1000){

    # FOR TWO COMPOSITIONS USING AITCHISON'S DISTANCE
# seals = EACH ROW IS A SEAL, THE GROUPS ARE rbind TOGETHER
# m = THE NUMBER OF GROUPS
# n1, n2 = THE NUMBER OF SEALS IN EACH GROUP RESPECTIVELY
# nperm = THE NUMBER OF PERMUTATIONS TO COMPUTE FOR THE P-VALUE
```

```
n <- n1 + n2

Forig <- npFtest4(seals, n1, n2, alpha)

# CALCULATING nperm RANDOM PERMUTATIONS AS DESCRIBED IN MARTI ANDERSON'S PAPER
# 1000 FOR 0.05 SIGNIFICANCE, 5000 FOR 0.01 SIGNIFICANCE

Fperm <- numeric(nperm)

for(i in 1:nperm){

ind <- c(1:nrow(seals))
perm <- sample(ind)
seals.perm <- seals[perm,]

Fperm[i] <- npFtest4(seals.perm, n1, n2, alpha)
}

# HOW MANY PERMUTATIONS WERE GREATER THAN THE ORIGINAL Forig
p.value <- sum(Fperm > Forig)/nperm
return(p.value)
}

# QUICK TRACE FUNCTION
tr <- function(mat){
trace <- sum(diag(mat))
return(trace)
}

twowaynpFtest <- function(seals, Xr, Xf, a, b){

# TESTS FOR FACTOR A EFFECT
```

```
# seals = MATRIX WHERE EACH ROW REPRESENTS A PROFILE OF A DIFFERENT INDIVIDUAL
# a = NUMBER OF FACTORS IN TREATMENT A
# b = NUMBER OF FACTORS IN TREATMENT B
# Xr = THE REDUCED DESIGN MATRIX EXCLUDING FACTOR A
      # Xf = THE FULL DESIGN MATRIX

N <- nrow(seals)

d <- matrix(0, nrow=N, ncol=N)
# USING AITCHISON'S DISTANCE MEASURE aDist, CALCULATING A DISTANCE MATRIX

for(i in 1:N){

for(j in 1:N){
d[i,j] <- aDist(seals[i,],seals[j,])
}
}

dsqr <- d^2
A <- (-1/2)*dsqr

Hr <- Xr%*%solve(t(Xr)%*%Xr)%*%t(Xr)
Hf <- Xf%*%solve(t(Xf)%*%Xf)%*%t(Xf)

ones <- matrix(1,N,1)
I <- diag(N)
g <- (I - (1/N)*ones%*%t(ones))
G <- g%*%A%*%g

SSA <- tr(Hf%*%G%*%Hf)-tr(Hr%*%G%*%Hr)
SSR <- tr((I-Hf)%*%G%*%(I-Hf))
SST <- tr(G)
```

```
F <- ((tr(Hf%*%G%*%Hf)-tr(Hr%*%G%*%Hr))/(a-1))/(tr((I-Hf)%*%G%*%(I-Hf))/(N-a-b+1))
return(F)
}

p.value2way <- function(seals, Xa, Xb, Xf, a, b, nperm=1000){

# P-VALUE FOR TWO WAY MANOVA
# seals = EACH ROW IS A SEAL, THE GROUPS ARE rbind TOGETHER
# nperm = THE NUMBER OF PERMUTATIONS TO COMPUTE FOR THE P-VALUE
# Xa = THE REDUCED DESIGN MATRIX EXCLUDING FACTOR A
# Xb = THE REDUCED DESIGN MATRIX EXCLUDING FACTOR B

Foriga <- twowaynpFtest(seals, Xa, Xf, a, b)
Forigb <- twowaynpFtest(seals, Xb, Xf, a, b)

# CALCULATING nperm RANDOM PERMUTATIONS AS DESCRIBED IN MARTI ANDERSON'S PAPER
# 1000 FOR 0.05 SIGNIFICANCE, 5000 FOR 0.01 SIGNIFICANCE

Fperma <- numeric(nperm)
Fpermb <- numeric(nperm)

for(i in 1:nperm){

ind <- c(1:nrow(seals))
perm <- sample(ind)
seals.perm <- seals[perm,]

Fperma[i] <- twowaynpFtest(seals.perm, Xa, Xf, a, b)
Fpermb[i] <- twowaynpFtest(seals.perm, Xb, Xf, a, b)
}

# HOW MANY PERMUTATIONS WERE GREATER THAN THE ORIGINAL Forig
```

```
p.valuea <- sum(Fperma > Foriga)/nperm
p.valueb <- sum(Fpermb > Forigb)/nperm
return(c(Foriga, p.valuea, Forigb, p.valueb))
}
```

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.

Anderson, M. (2001). A new method for non-parametric multivariate analysis of variance. *Austral Ecology* 26 32–46.

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¡ i¿ heterodontus portusjacksoni¡/i¿. *Journal of Experimental Marine Biology and Ecology* 448 10–18.

Brander, K. (2010). Impacts of climate change on fisheries. *Journal of Marine Systems* 79 389–402.

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.

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.

Fisk, A., Tittlemier, S., Pranschke, J. & Norstrom, R. (2002). Using anthropogenic contaminants and stable isotopes to assess the feeding ecology of greenland sharks. *Ecology* 83 2162–2172.

Graham, N., McClanahan, T., MacNeil, M., Wilson, S., Polunin, N., Jennings, S., Chabanet, P., Clark, S., Spalding, M. & Letourneur, Y. (2008). Climate warming, marine protected areas and the ocean-scale integrity of coral reef ecosystems. *PLoS one* 3 e3039.

Greenacre, M. (2010). Log-ratio analysis is a limiting case of correspondence analysis. *Mathematical Geosciences* 42 129–134.

Greenacre, M. (2011). Measuring subcompositional incoherence. *Mathematical Geosciences* 43 681–693.

Hovde, S., Albert, O. & Nilssen, E. (2002). Spatial, seasonal and ontogenetic variation in diet of northeast arctic greenland halibut (reinhardtius hippoglossoides). *ICES Journal of Marine Science: Journal du Conseil* 59 421–437.

Iverson, S., Field, C., Bowen, W. & Blanchard, W. (2004). Quantitative fatty acid signature analysis: a new method of estimating predator diets. *Ecological Monographs* 74 211–235.

Leclerc, L., Lydersen, C., Haug, T., Glover, K., Fisk, A. & Kovacs, K. (2011). Greenland sharks (somniosus microcephalus) scavenge offal from minke (balaenoptera acutorostrata) whaling operations in svalbard (norway). *Polar Research* 30.

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.

MacNeil, M., McMeans, B., Hussey, N., Vecsei, P., Svavarsson, J., Kovacs, K., Lydersen, C., Treble, M., Skomal, G. & Ramsey, M. (2012). Biology of the greenland shark somniosus microcephalus. *Journal of Fish Biology* 80 991–1018.

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. 526–531.

Martín-Fernández, J. & Thió-Henestrosa, S. (2006). Rounded zeros: some practical aspects for compositional data. *Geological Society, London, Special Publications* 264 191–201.

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.

McMeans, B., Arts, M., Lydersen, C., Kovacs, K., Hop, H., Falk-Petersen, S. & Fisk, A. (2013). The role of greenland sharks (somniosus microcephalus) in an arctic ecosystem: assessed via stable isotopes and fatty acids. *Marine Biology* 1–16.

Overpeck, J., Sturm, M., Francis, J., Perovich, D., Serreze, M., Benner, R., Carmack, E., Chapin, F., Gerlach, S., Hamilton, L. et al. (2005). Arctic system on trajectory to new, seasonally ice-free state. *Eos, Transactions American Geophysical Union* 86 309–313.

Pawlowsky-Glahn, V. & Buccianti, A. (2011). *Compositional data analysis: Theory and applications*. John Wiley & Sons.

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. ("In Review"). Testing for a change in diet using fatty acid signatures. *Ecological and Environmental Statistics* .