An approach to measure distance between compositional diet estimates containing essential zeros

Connie Stewart

Journal of Applied Statistics, Volume 44, Issue 7

DOI: https://doi.org/10.1080/02664763.2016.1193846
Publisher: Taylor & Francis
Published Article URL: https://doi.org/10.1080/02664763.2016.1193846
Published Issue URL: https://www.tandfonline.com/toc/cjas20/44/7?nav=tocList

To appear in the *Journal of Applied Statistics*
Vol. 00, No. 00, Month 20XX, 1–15

An approach to measure distance between compositional diet estimates containing essential zeros

Connie Stewart

(*Received 00 Month 20XX; accepted 00 Month 20XX*)

For many applications involving compositional data it is necessary to establish a valid measure of distance, yet when essential zeros are present traditional distance measures are problematic. In quantitative fatty acid signature analysis (QFASA), compositional diet estimates are produced that often contain many zeros. In order to test for a difference in diet between two populations of predators using the QFASA diet estimates, a legitimate measure of distance for use in the test statistic is necessary. Since ecologists using QFASA must first select the potential species of prey in the predator's diet, the chosen measure of distance should be such that the distance between samples does not decrease as the number of species considered increases, a property known in general as subcompositional coherence.

In this paper we compare three measures of distance for compositional data capable of handling zeros, but not satisfying some of the well-accepted principles of compositional data analysis. For compositional diet estimates, the most relevant of these is the property of subcompositionally coherence and we show that this property may be approximately satisfied. Based on the results of a simulation study and an application to real-life QFASA diet estimates of grey seals, we recommend the chi-square measure of distance.

Keywords: Chi-square distance; Compositional data; Essential zeros; QFASA; Subcompositional coherence.

1. Introduction

Measuring distance between two compositions in a valid manner is important in many applications involving compositional data. While in this paper we require a measure of distance between two compositional data sets for use in a multivariate permutation test, measures of distance are also the basis of clustering methods ([12],[20]) and of recently developed nonparametric multivariate analysis of variance procedures ([14]), for example. Yet when zeros are present, traditional distance measures either cannot be applied directly (due to the presence of logarithms, ratios, etc...in the distance definition) or do not satisfy certain properties of compositional data analysis considered to be fundamental in [1] and [4]. While [20] presents a thorough discussion of available measures of dissimilarity between compositions, they do not deal with what they recognize as the "important practical problem" of compositions with zeros, and instead refer readers to recent work in this area. We note that the specified papers focus on the treatment of zeros of the *rounded* type and not *essential zeros* which are present in our application.

In this paper we compare three measures of distance for compositional data that are capable of handling zeros of any type, but do not satisfy the principle of subcompositional coherence (defined formally in Section 2) which is relevant for our application of interest. Following the basic ideas in [7], we attempt to measure the subcompositional incoherence

Department of Computer Science and Applied Statistics, University of New Brunswick Saint John, Saint John, N.B., Canada, E2L 4L5. Email: cstewart@unb.ca

of the measures in order to determine if the measures may, from a practical point of view, satisfy the principle approximately. In [7], however, subcompositional coherence was measured in a different context since the author was interested in distances between components while we are concerned with distances between individuals. One consequence of this is that we cannot use the author's measure of incoherence, or "stress", directly, hence we propose and examine alternate measures of "stress". Furthermore, we have broadened the scope of work in this area to include three measures of distance and, importantly, a study of the effect of zeros on the degree of subcompositional incoherence of these measures.

In our motivational application, the data consists of diet estimates containing the estimated proportion of each species of prey in a predator's diet determined from an approach called quantitative fatty acid signature analysis (QFASA) in which predator fatty acid (FA) signatures are matched to their prey signatures ([8]). Since its introduction, QFASA has become an increasingly popular method of diet estimation, particularly for marine species. Here we are interested in the basic but important ecological problem of determining whether the diets of two groups of predators are different based on the QFASA estimates. Because of the small sample sizes that are often associated with QFASA, we prefer to use a nonparametric approach which, in turn, requires a legitimate measure of distance between the samples. Since the diet estimates often contain many zeros and these zeros represent an estimated absence of a species from the diet, the recommended and usual approach is to treat the zeros as essential zeros rather than replacing them with an arbitrary amount ([17]). While strategies for replacing *rounded* zeros by appropriate values have been developed (see [13, 15], for example), these are not recommended for essential zeros. In addition to QFASA, other diet estimation methods result in compositional data with zeros and methods presented here may be easily extended to them.

We note that [18] examined the type one error and power of multivariate nonparametric tests designed to test for a difference in the compositional FA signatures of the predator themselves which are known to be related to diet. As the authors point out in their paper however, a difference in FA signatures does not always translate to a difference in diet and using the QFASA diet estimates instead allows for a more direct test. Working at the FA signature level is somewhat simpler because zeros occur in the data to a much lesser extent and, in practice, are often treated as rounded zeros. Ultimately we would like to apply the methods in [18] to the diet estimates with a measure of distance determined to be appropriate for compositional data with many essential zeros.

We begin in Section 2 with pertinent definitions related to measures of compositional difference. A simulation study is described in Section 3 in which the performance of the measures of distance is assessed in a variety scenarios. In Section 4, the measures of distance are utilized in a real-life setting involving QFASA diet estimates of grey seals inhabiting the east coast of Canada. We use the multivariate permutation test proposed by [18] for detecting changes in the FA signatures (but applied to the diet estimates) along with our examined measures of distance to determine if there is a difference in the diets of male and female grey seals prior to breeding season.

2. Measures of Distance between Compositions

When the data is compositional, such as in our QFASA application (that is, where elements are non-negative and sum to one), it has been argued that a distance measure should satisfy certain fundamental principles of compositional data analysis. In [4] these are specified to be scale invariance (distance$(k\mathbf{y}, K\mathbf{y}^*) =$ distance$(\mathbf{y}, \mathbf{y}^*)$ for all positive

constants k and K and compositions $\mathbf{y}$ and $\mathbf{y}^*$), permutation invariance (re-ordering the parts should not change conclusions) and, in reference to the analysis of subcompositions, subcompositional coherence. Note that a subcomposition is a vector consisting of a subset of parts that has been "closed" so that it sums to one. In [4], two criteria for subcompositional coherence of a distance measure are provided: 1) scale invariance should hold for any subcomposition and 2) the distance between two compositions should be larger or equal to the distance between corresponding subcompositions. Strictly speaking, the latter criteria is called subcompositional dominance though here we use the terms coherence and dominance interchangeably.

While intuitively subcompositional coherence is a sensible requirement, in a practical setting there is sometimes a "price to be paid" for attempting to adhere stringently to the above mentioned properties, as pointed out in [21]. For example, as discussed below, requiring exact subcompositional coherence makes dealing with zeros especially challenging. In the context of our problem, subcompositional dominance essentially ensures that the distance between two samples of QFASA diet estimates based on say 30 potential species of prey is at least as great as the distance between the samples using a reduced prey database containing, for example, 5 species. Or, in other words, adding species to the prey database used to estimate the diet should not produce samples of diet estimates that are closer together and we feel this property is therefore both relevant and important.

In many applications perturbation invariance (equivalent to translation invariance in real space) is also a logical requirement as it guarantees that the distance is the same regardless of the units of measurement used. (See [20] for a more detailed discussion concerning distance measures and this property). Because QFASA applications do not involve changing units, we do not view this property as being particularly relevant for our problem of interest and it is not discussed further in this work.

In relation to hierarchical clustering methods for compositional data, [12] examined a variety of distance measures. Only Aitchison's distance measure (and another based on Aitchison's methods) satisfied all of the aforementioned principles including subcompositional coherence. Aitchison's distance between two compositions $\mathbf{Y}_1$ and $\mathbf{Y}_2$ of length p (as defined in [12]) is as follows:

$$\mathrm{AIT}(\mathbf{Y}_1, \mathbf{Y}_2) = \left(\sum_{j=1}^{p} \{\log[Y_{1j}/g(\mathbf{Y}_1)] - \log[Y_{2j}/g(\mathbf{Y}_2)]\}^2 \right)^{1/2} \tag{1}$$

where $g(\mathbf{Y}) = (Y_1 \cdots Y_p)^{\frac{1}{p}}$ represents the geometric mean. When essential zeros are present, such as in the QFASA application, Aitchison's distance measure is clearly not useful.

The angular and (crude) Mahalanobis distance measures are also defined in [12] and these are given as scale as well as permutation invariant distance measures, capable of handling zeros, but are not subcompositionally dominant. In [12], the angular (ANG) distance between two compositions $\mathbf{Y}_1$ and $\mathbf{Y}_2$ was defined as

$$\mathrm{ANG}(\mathbf{Y}_1, \mathbf{Y}_2) = \arccos\left(\sum_{j=1}^{p} \sqrt{\frac{Y_{1j}^2}{\sum Y_{1j}^2}} \sqrt{\frac{Y_{2j}^2}{\sum Y_{2j}^2}} \right), \tag{2}$$

and the Mahalanobis (MAH) distance as

$$\mathrm{MAH}(\mathbf{Y}_1, \mathbf{Y}_2) = \left[(\mathbf{Y}_1 - \mathbf{Y}_2)' \mathbf{K}^+ (\mathbf{Y}_1 - \mathbf{Y}_2) \right]^{1/2} \tag{3}$$

where the matrix $\mathbf{K}^+$ denotes the Moore-Penrose pseudo-inverse of the covariance matrix $\mathbf{K}$ of a compositional data set. Since in our application we have two compositional data sets (one for each population of predators), we have chosen to let $\mathbf{K}$ be the pooled sample covariance matrix and, defined in this manner, this distance measure is no longer scale invariant in general. Another conceivable drawback of this distance measure is that it is not well-defined for arbitrary compositions because samples are needed in order to calculate $\mathbf{K}^+$. Being sample dependent could also, for example, have implications on conclusions concerning coherence. In spite of these issues, we have chosen to include the MAH distance measure in our comparison study given that it does allow for problematic essential zeros. Moreover, if the measure of distance is applied to previously normalized data and if samples of compositions are available (which is the case in our motivating application), then these potential issues should be relatively minor.

Our third measure of distance of interest was recently defined in [18] and was based on the chi-square distance measure given in [6] for use in correspondence analysis. According to [9], correspondence analysis is widely accepted in the ecology community. We define the chi-square (CS) measure of distance between $\mathbf{Y}_1$ and $\mathbf{Y}_2$ as:

$$\mathrm{CS}(\mathbf{Y}_1, \mathbf{Y}_2) = \sqrt{2p}\left(\sum_{j=1}^{p} r_j\right)^{1/2}, \tag{4}$$

where

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

The CS distance is scale and permutation invariant, allows for zeros, but is not subcompositionally dominant. Another salient point is that the ANG and MAH distance measures are known to be true metrics and in particular, satisfy the triangle inequality. Based on simulations our conjecture is that the inequality does in fact hold for the CS measure as well, but this remains to be shown formally. In any case, unlike the property of subcompositional coherence, it is not obvious that this property is of practical importance for our motivating application.

It is not at all clear how to devise a distance measure that both satisfies the principle of subcompositional coherence and is capable of handling essential zeros, since the principle implicitly requires that ratios of components be compared where the denominator could be zero in applications involving essential zeros. In our search for a valid distance measure capable of handling essential zeros, we have therefore opted to use an alternative approach, based on ideas presented in [7], and instead of requiring absolute subcompositional coherence, we examine otherwise valid measures of distance for which essential zeros are not an issue (that is, Equations 2-4) and attempt to measure the subcompositional incoherence of the measures using measures of stress.

It should perhaps be mentioned that in [18] a more general definition of the CS distance is provided which includes a power transformation parameter, γ. With this definition and a specific value for γ, the data is power transformed (by γ) before Equation 4 is used to calculate distance. When there are no zeros in the compositional data vectors of interest, it can be shown that as γ tends to 0, the CS distance converges to the AIT distance, essentially as a result of the well-known Box-Cox transformation ([5, 6]). In [18], the parameter γ was chosen by considering decreasing values of γ until subcompositional coherence is essentially achieved. More specifically, given two compositional data sets

and γ, the authors measured subcompositional incoherence by first computing a distance matrix, say $\mathbf{D}$, consisting of all pairwise distances between full compositions and then, for a given 2-part subcomposition (since 2-part subcompositions were found in [7] to represent a worst case scenario), computed a similar distance matrix, say $\mathbf{S}$. The two matrices were then compared and the stress was calculated as the proportion of entries in $\mathbf{D}$ for which $d_{ij} < s_{ij}$. The power transformation for which the average stress over all 2-part subcompositions was near zero was selected.

Having to determine γ prior to using the CS distance, however, presents an added complexity from a practical perspective. Recall that [7] measured coherence in a different manner and it is not known in the context of our problem if 2-part subcompositions yield the largest stress when the CS distance is used. In fact, using a different size subcomposition to compute the stress may lead to a different choice of γ and possibly inaccurate conclusions. Furthermore, exploratory work not presented here suggested that when many essential zeros were present, the CS distance becomes unstable when γ is close to zero and that for this application incoherence was at a minimmum when $\gamma = 1$. Lastly, the results of our simulation study imply that the additional parameter is not needed in order for the CS measure of distance to be, practically speaking, subcompositional coherent. For these reasons, we have opted to drop the power transformation parameter (or equivalently set $\gamma = 1$) in our definition of the CS distance.

3. Simulation Study

3.1 *QFASA*

To carry out a simulation study relevant to QFASA, we require the ability to generate samples of compositional data representing data that would arise in practice. We accomplish this by simulating QFASA diet estimates which contain estimates of the proportion of individual species in a predator's diet.

QFASA was developed by [8] and is essentially based on the concept that for some predators, such as seals, the FAs of the prey consumed are deposited in the adipose tissue of the predator with little modification. In simplistic terms, given a predator FA signature and a sample of prey FA signatures from each species of prey potentially in the predator's diet (the latter sample forms a prey database), the QFASA estimate is then the weights determined by minimizing the distance (such as AIT distance) between the predator FA signature and a weighted mixture of the mean prey FA signatures. Additional biological factors need to be considered in order to improve the accuracy of the estimates, and these are discussed in Section 4 where real-life samples of grey seal FA signatures are analyzed.

Given a prey database we can then generate pseudo-predators and obtain the QFASA diet estimates of the pseudo-predators. Pseudo-predators were first introduced in [8] and a modified algorithm was developed in [17]. A pseudo-predator is created by sampling with replacement from a prey database with probability weights given by a selected 'true' diet. The prey database used in our simulation study was collected along the Scotian Shelf off eastern Canada and contains 28 species of fish (2110 FA signatures in total) which could potentially be part of the diet of a seal inhabiting the east coast of Canada ([3]). While FA signatures (pseudo-predators) of dimension 65 are generated, only a subset of 40 FAs known to arise from diet and/or biosynthesis are used in the simulations to obtain the diet estimates.

Our simulations are carried out by generating two independent samples of FA signatures each with a selected true diet. The diet of each pseudo-predator is estimated via QFASA yielding two corresponding samples of diet estimates. Note that the true diet is

a vector of proportions with component i corresponding to the true desired proportion of species i in the pseudo-predator's diet. If the true diet contains many zeros, then the diet estimates will also tend to contain many zeros. In Section 3.3 where we investigate the effect of essential zeros, the pseudo-predators themselves (or equivalently FA signatures) will be useful on their own since they are also compositional but do not contain any zeros and can therefore serve as a baseline for comparison purposes.

3.2 *Measuring Subcompositional Incoherence*

For two matrices of compositional data of dimension $n_1 \times p$ and $n_2 \times p$, we measure subcompositional incoherence of a distance measure by comparing the distance matrix, $\mathbf{D}$, consisting of all pairwise distances between full compositions, with a similar distance matrix $\mathbf{S}_c$ calculated from c-part subcompositions.

To compare $\mathbf{D}$ and $\mathbf{S}_c$, we define two measures of stress. The first measure, stress_1, is simply the measure of stress used in [18], discussed previously in Section 2, and is defined more formally as:

$$\text{stress}_1(\mathbf{D}, \mathbf{S}_c) = 1 - \frac{1}{n_1 n_2} \sum_i \sum_j I(d_{ij}, s_{cij}), \tag{5}$$

where $I(d_{ij}, s_{cij}) = 1$, if $d_{ij} \geq s_{cij}$ and 0 otherwise.

This simple measure of stress takes into account only whether $d_{ij} < s_{ij}$ and not the magnitude of the difference. The need for an alternate measure arose from initial findings in which we observed that when $d_{ij} > s_{cij}$, it was often by an insignificant amount. Hence we also measure stress by

$$\text{stress}_2(D, S_c) = \sqrt{\frac{\sum_i \sum_j (h_{ij}^2)^2}{\sum_i \sum_j d_{ij}^4}} \tag{6}$$

where $h_{ij}^2 = d_{ij}^2 - s_{cij}^2$ if $d_{ij} < s_{ij}$ and 0 otherwise.

Equation 6 is a modified version (since it takes into account distances for which $d_{ij} < s_{cij}$ only) of the recommended measure of discrepancy in [10] (Chapter 12). Although working in a different context, this measure of stress was introduced by [19] and, according to [10], "is becoming the preferred criterion". It should perhaps also be mentioned that if we modified the stress formula defined in [7] in an analogous manner (i.e. to only take into account distances for which $d_{ij} < s_{cij}$), we would obtain a measure of stress similar to stress_2 . In fact, although not shown, the behavior of this measure of stress was analogous to stress_2, particularly for the MAH and CS measures of distance in simulations. While it is possible to devise other appropriate measures of stress for this problem, using two distinct measures of stress (namely 5 and 6) should adequately allow conclusions to be made regarding incoherence among the measures of distance.

The reported stress is actually the average stress over many subcompositions of size c. Specifically, since for some values of c, the number of possible subcompositions is very large (for example, $\binom{28}{10} = 13,123,110$), we chose to simply average the stress over $\min\left(100, \binom{28}{c}\right)$ subcompositions for a given choice of c, rather than averaging over all possible subcompositions. Note that, particularly when c is small, the sample of subcompositions often contain one or more c-part sub vectors with components all zeros when the true diet contains many zeros. When this occurs in the simulations, the corresponding observations are removed from the sample with the rationale being that a subcomposi-

Table 1. Diets used in simulations for assessing measures of distance.

Species	Diet A	Diet B
Northern Sandlance	35	50
Redfish	0	0
Capelin	0	0
Atlantic Cod	30	15
Silver Hake	15	5
American Plaice	0	0
Yellowtail Flounder	10	20
Longhorn Sculpin	0	0
Other	10	10
TOTAL	100	100

tion consisting of all-zeros is unlikely to appear in practice. If too few observations (here we used ≤ 2) remain after the zero subvectors are removed, another c components are randomly chosen.

3.3 *Results*

We carried out simulations to assess 1) the overall subcompositional coherence of the three measures of distance for typical samples of diet estimates and for various subcomposition sizes and 2) the effect of essential zeros on subcompositional coherence.

With respect to our first objective, we made use of two of the diets examined in [18] known to be representative of the diet of grey seals on the east coast of Canada. The distance between the diets is such that samples generated from these diets typically differed significantly. We call these diets "Diet A" and "Diet B" and they are specified in Table 1. Note that "Other" refers to noise, or random sampling from the remaining species in the prey database. Based on the true diets, there may be potentially many zeros in the diet estimates. From a practical point of view, it may be helpful to know if there is a particular value of c for which the stress is generally higher. We examined the stress for $c = 2, 5, 10, 15, 20, 25, 27$ (recall that $p = 28$ since there are 28 potential species of fish in a seal's diet) and at $n = n_1 = n_2 = 10, 25$. These smaller sample sizes were chosen in accordance with the real-life data sample sizes often observed in practice including the data examined in Section 4. (Because of the potential for all-zero subcompositions when c is small, the actual sample sizes may be smaller than 10 and 25.)

In summary, for each of M iterations, two samples of pseudo-predators (one with Diet A and the other with Diet B) and corresponding diet estimates are generated and the average stress (either stress_1 or stress_2) is computed for each value of c. The stress is then again averaged over all M samples, where we chose to use $M = 500$ since results are computationally intensive to obtain.

Results from this part of the simulation study are displayed graphically in Figure 1. For both the ANG and CS measures of distance, the sample size does not appear to influence the stress, but this is not the case for the MAH distance. Although not shown in the graph, when $n = 10$, the average stress_2 for the MAH distance was approximately 1609.2, 47.1, and 1.1 for subcomposition sizes 20, 25, and 27 respectively. We surmise that the issue is related to computation of the pseudo inverse matrix at the smaller sample size. By $n = 25$, the pseudo inverse matrix appears to have become more stable producing reasonable results. This issue of unusually large stress values is not present for stress_1 nor for smaller subcomposition sizes.

Based on the figures, both measures of stress behave similarly with respect to how

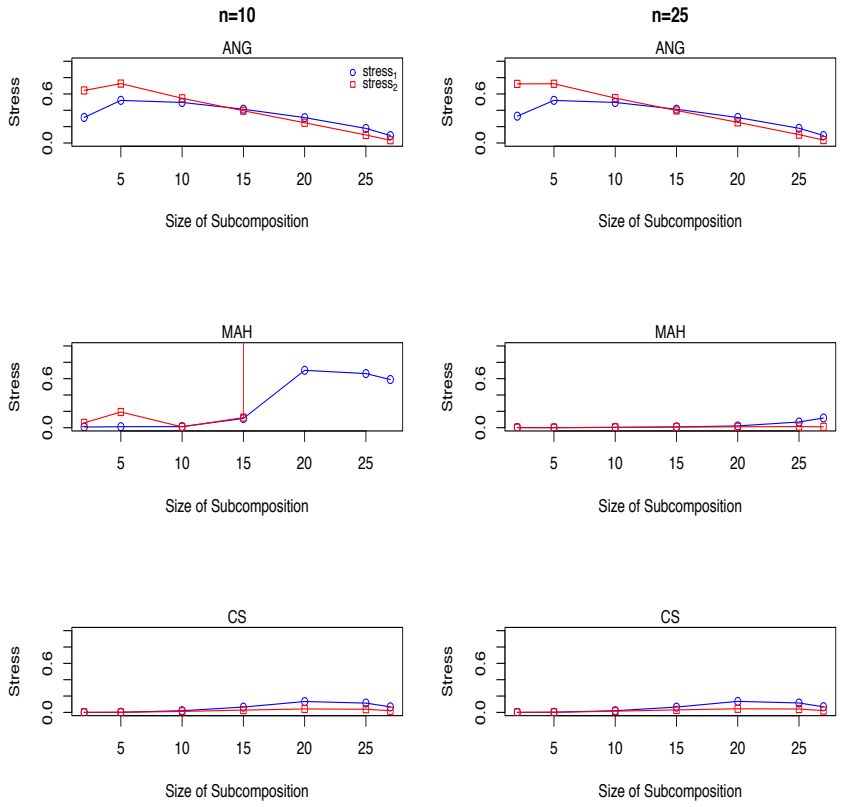

Figure 1. Stress associated with angular, crude Mahalanobis and chi-square measures of distance for typical samples of diet estimates ($n = 10$ and $n = 25$). Note that for $n = 10$, MAH distance values greater than one are not shown.

they vary with subcomposition size. For the ANG distance, the stress (both measures) tends to peak around $c = 5$ and for the MAH and CS measures of distance, the stress is generally higher for the larger subcomposition sizes. It is interesting to note that while the extent of the subcompositional incoherence clearly varies with the size of the subcomposition, 2-part subcompositions do not appear to represent the worst case scenario for this application.

In terms of the magnitude of the stress and practicality of the measures, the CS and MAH (at $n = 25$ only) yielded reasonably small values of stress for all of the subcomposition sizes examined. This is particularly the case if stress_2 is considered. We speculate that the difference between stress_1 and stress_2, with stress_1 giving larger values, is due to the distance between two full compositions often being only marginally smaller than the distance between the corresponding subcompositions.

The second goal of our simulation study was to gain a better understanding of the effect of essential zeros on the subcompositional incoherence of the measures of distance. Since the simulated pseudo-predators (of dimension 40) do not contain any zeros, we first computed the stress for samples of pseudo-predators (averaged over $M = 500$ samples) associated with the measures of distance for three subcomposition sizes: 1) small ($c = 5$), 2) medium ($c = 20$) and 3) large ($c = 35$). As explained below, it is typically not feasible to generate diet estimates containing only few or no zeros (unless the diet estimates are adjusted so we have instead used the FA signatures to accomplish this. We used Diet A and Diet B in Table 1 and examined only a sample size of 25 since in our previous results

Table 2. One set of randomly selected diets used in simulations for assessing the effect of zeros.

	I		II		III	
Species	Diet A	Diet B	Diet A	Diet B	Diet A	Diet B
Cod	10	24	29	0	52	0
Haddock	0	13	0	27	0	68
Plaice	23	0	0	15	0	0
Pollock	30	19	0	24	0	0
Sandlance	17	16	39	0	48	0
Silverhake	0	12	22	0	0	0
Winter Flounder	18	0	10	0	0	0
Yellowtail	2	16	0	34	0	32
% Zeros in Diet	25	25	50	50	75	75

a sample size of $n = 10$ was problematic for the MAH distance measure for the larger subcomposition sizes, but otherwise results were almost identical for both sample sizes.

To examine the effect of essential zeros on the stress, we used a reduced prey base containing 8 important species, as described in [17]. This allowed us to vary the amount of zeros in the true diets and produce diet estimates that are generally closer to the true diet than would be expected if the larger prey data base was used. However, for a particular choice of true diet with a specified number of zero components, it will still often be the case that the zeros in the diet estimate of a generated pseudo-predator do not coincide with the zeros in the true diet. To help keep the percentage of zeros in the diet estimates similar to the true diet, we replaced components in the diet estimates by zero if the corresponding components in the true diet were also zero, and renormalized the estimates.

We examined three settings: 1) Two zero components in the true diet (or 25% of components being zero), 2) four zero components in the true diet (or 50% of components being zero) and 3) six zero components in the true diet (or 75% of components being zero). For each of these three settings, each true diet (that is, Diet A and Diet B) was chosen by first randomly selecting which components would be zero and then simulating the non-zero components from a uniform $(0, 1)$ distribution. The resulting normalized vectors were used as the true diets. For instance, Table 2 displays the true diets for one random selection of true diets. For a given setting, $M = 500$ samples of diet estimates are generated from Diet A and also from Diet B and the stress is computed as before for all three measures of distance using subcomposition sizes 3, 6 and 7.

Figures 2 and 3 display the stress results associated with the three measures of distance. When the data (consisting of FA signatures) contain no zeros (Figure 2), the stress associated with the MAH and CS measures of distance are similar and consistently smaller than for the ANG distance measure. With respect to the effect of the percentage of zeros in the data (Figure 3) on subcompositional coherence, using our measures of stress, it appears that the ANG distance measure is more subcompositionally coherent when there are more zeros in the data. This also appears to be true for the CS distance but to a much lesser extent. Overall the stress (either measure) is small for the MAH (at least for $n = 25$) and the CS measures of distance irrespective of the percentage of zeros in the data, suggesting that these two measures may be sufficiently subcompositionally coherent from a practical perspective.

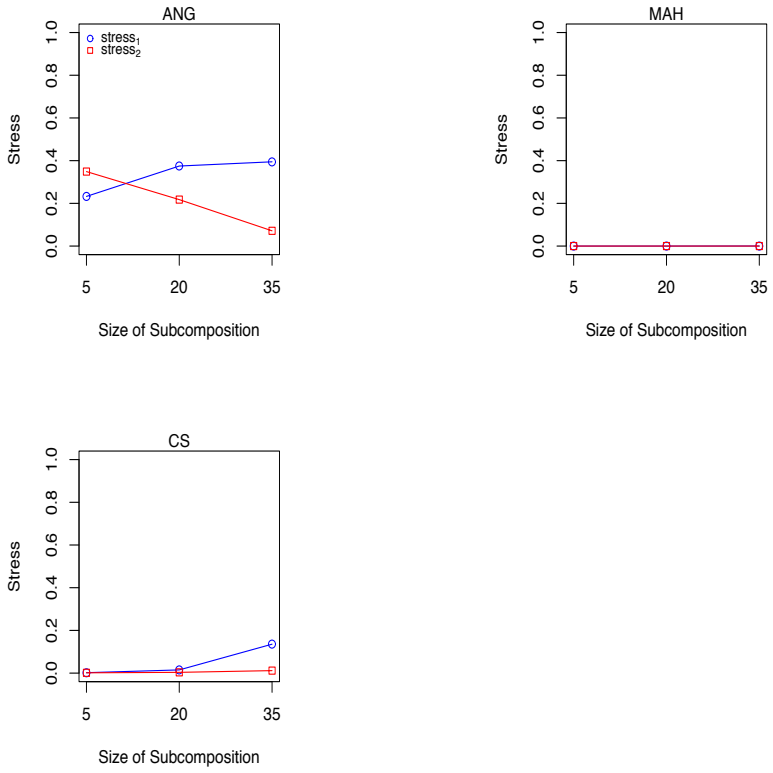

Figure 2. Stress associated with the angular, crude Mahalanobis and chi-squared measures of distance for sample sizes of $n = 25$ and for three subcomposition sizes when the data does not contain any zeros.

4. Real-life Application

Recall that our study of measures of distance was motivated by the problem of using QFASA diet estimates to determine whether the diet of groups of predators differed. Variation in diet within a species and the reason for this variation (such as gender, season, year, etc...) is important to biologists studying the role of the species in the ecosystem and to ecosystem managers ([2].) Here we are interested in comparing the diets of male and female grey seals living on the east coast of Canada prior to the breeding season (December-January) of 2011. Analyses based on QFASA estimates of diet present some statistical challenges as they are compositional, contain many essential zeros and the sample sizes are small compared to the dimension of the diet estimate vectors. To manage these issues, we apply the multivariate nonparametric test considered in [18] with a test statistic defined as the sum of all distances between the samples where the distance is either ANG, MAH or CS. We now describe the analysis in more detail beginning with a description of the data.

4.1 *Data*

As part of a larger, ongoing study which is investigating spatial and temporal variation in grey seal foraging behaviours and diet ([2] and WD Bowen, personal communication), full-depth blubber biopsies were collected early in the breeding season from 6 male and 26

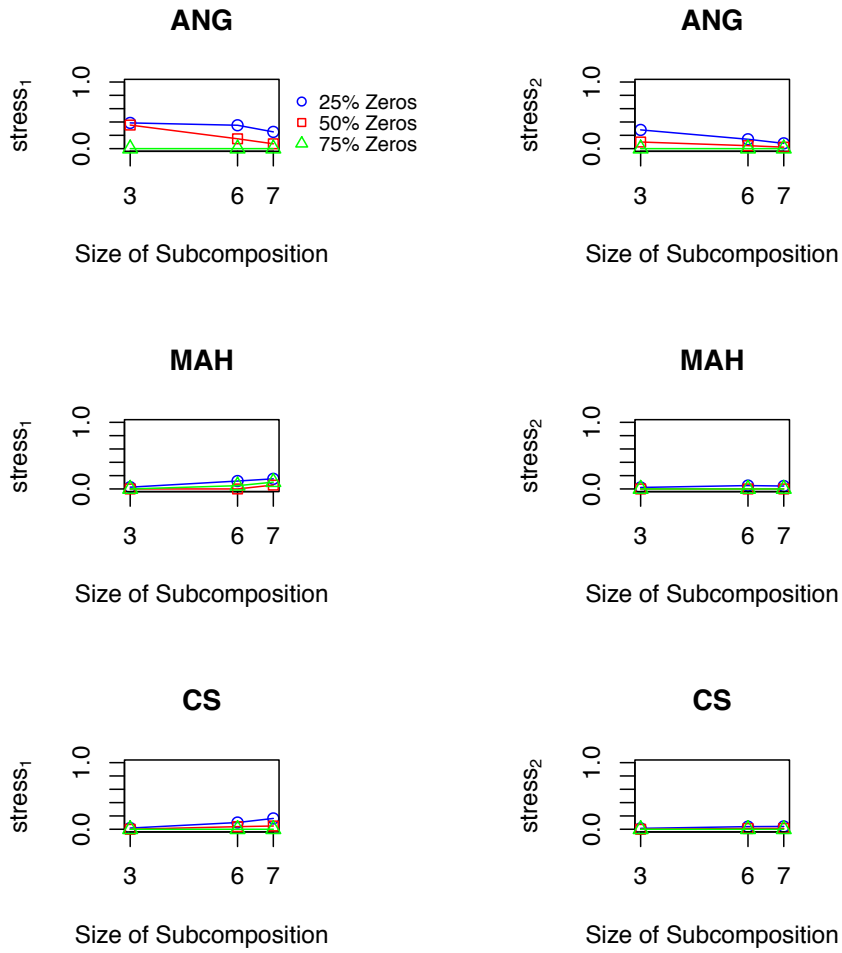

Figure 3. Stress associated with angular, crude Mahalanobis and chi-squared measures of distance for a sample sizes of $n = 25$ with generated diet estimates containing varying amounts of zeros, for three different sizes of subcompositions.

female grey seals on Sable Island (43°55′N, 60°00′W) following [11]. Lipid was extracted from each sample and the FA composition determined following the methods described in [2]. For this analysis, only a subset of 38 of the quantified FAs was used to obtain the diet estimates and these are available from the author. Furthermore, it is known that some FAs in the prey are not deposited exactly in the seal's adipose tissue and may always be higher or lower in the seal than in the prey. To handle this, [8] introduced calibration factors and we accordingly also calibrate the FAs before estimating the diet. A discussion of the calibration factors used in our analysis may be found in [16].

Recall that QFASA also requires a comprehensive prey database. Our QFASA estimates are determined from a modified version of the database previously discussed in Section 3. Biologists Shelley Lang (Dalhousie University) and Don Bowen (Department of Fisheries and Oceans) re-examined and subsequently updated this prey database in order to improve the accuracy of QFASA. We used the prey database (unpublished) that they recommend for seals sampled in the winter season. The 26 species and corresponding sample sizes in this prey database are provided in Table 3. Note that because of the often large within species variability in the prey database (due in large part to seasonal differences in the prey signatures), three species, namely American plaice, Capelin and

Table 3. Species and sample sizes used in QFASA.

Species	Sample Size
American lobster	21
American plaice	134
Atlantic butterfish	26
Atlantic cod	109
Atlantic herring	121
Atlantic mackerel	32
Capelin	48
Winter flounder	50
Witch flounder	24
Yellowtail flounder	156
Gaspereau	70
Haddock	115
Silver hake	58
White hake	80
Longhorn sculpin	25
Northern sandlance	148
Northern short squid	35
Northern shrimp	110
Ocean pout	31
Pollock	53
Redfish	54
Sea raven	71
Smooth skate	33
Thorny skate	83
Winter skate	40
Snake blenny	18

Table 4. Multivariate permutation test p-values comparing the male and female grey seals prior to breeding season of 2011 for the three measures of distance.

Measure of Distance	P-value
ANG	0.021
MAH	0.033
CS	0.011

Pollock were subdivided into two prey types (for example, Plaice 1 and Plaice 2) but their contributions to the diet were combined in the final diet estimate.

4.2 *Results*

QFASA diet estimates are shown in Figure 4 for the male and female grey seals. Note that for the male seals, approximately 42% of the species in the average diet estimate were zero and 58% for the female seals. From the plot, it appears that a large part of the grey seals' diet prior to the breeding season of 2011 was comprised of Atlantic herring, Capelin and Redfish which together account for approximately 82% of the females' average diet and 59% of the males' average diet. Our estimates suggest that the male seals are eating a wider variety of species of fish than the female seals. The three p-values associated with the multivariate permutation test were all small and not drastically different, and we conclude that there is a significant difference between the male and female diets. Gender

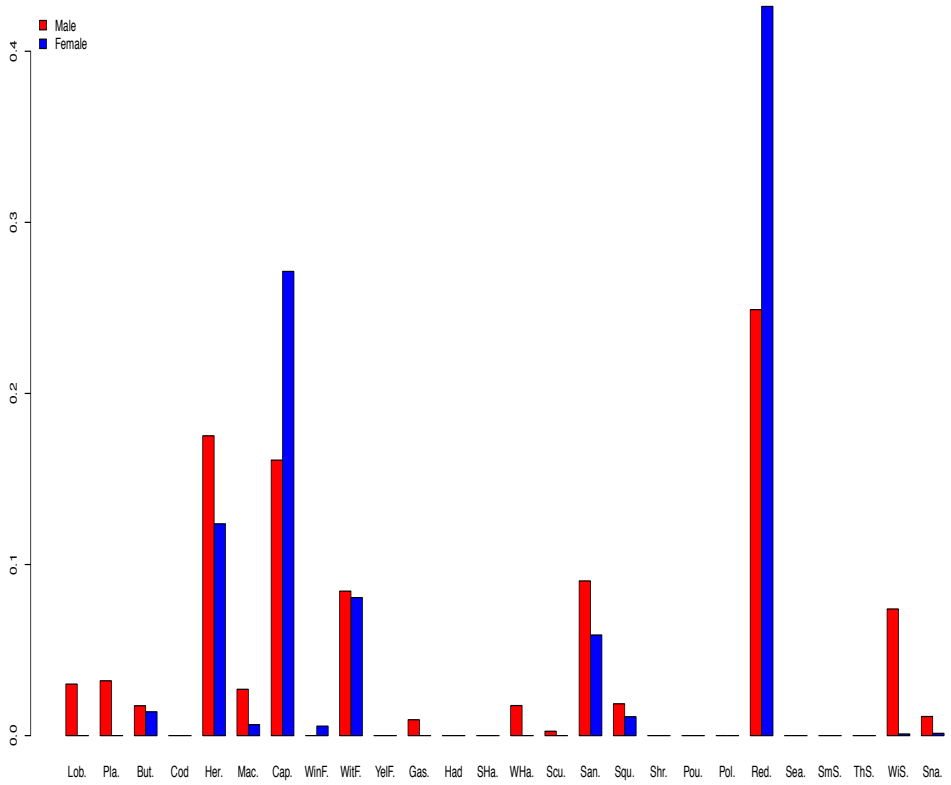

Figure 4. QFASA diet estimates of male and female grey seals prior to breeding season of 2011.

appears to be playing a role in the prey consumption of these grey seals.

5. Discussion

Measures of distance underpin many of the statistical methods available for multivariate data. Unfortunately, when the data of interest is compositional with essential zeros, there is not to our knowledge a distance measure available that satisfies the principle of subcompositional coherence and, in particular, subcompositional dominance, a property that we consider to be important in our compositional data application. Our overall goal was to determine whether any of the three measures of distance, all capable of handling compositional data with essential zeros, were approximately subcompositionally coherent, particularly when the data consists of QFASA diet estimates. Although we followed the usual approach of treating the zeros in the QFASA diet estimates as essential zeros, future work comparing QFASA results for this and other problems when zeros are instead simply modified by a small amount might be of interest to determine if, from a practical perspective, it is worthwhile to sacrifice exact subcompositional coherence.

Of the measures considered, we recommend the CS measure of distance since its associated stress (measured in two ways and argued to be related to the subcompositional coherence of the measure) was low for typical samples of diet estimates as well as for samples containing relatively large amounts of zeros. The ANG distance measure tended to yield larger stress values suggesting that it is less subcompositionally coherent. While

the MAH distance measure performed very well for samples of size 25 in terms of yielding small values of stress, it was problematic when both sample sizes were 10 which may preclude its use in some QFASA problems. The issue appeared to be with the computation of the pseudo inverse matrix. Being sample dependent and requiring a covariance matrix may be viewed as a disadvantage. For our application we needed the additional assumption of equal covariance matrices for the two populations of predator diet estimates which may not always be a reasonable assumption. Furthermore, unlike the scale invariant CS and ANG measures of distance, the MAH distance measure is not, in general, scale invariant.

Note that none of the measures presented here are perturbation invariant nor has the CS measure of distance been shown to satisfy the triangle inequality and, consequently, the measures may not be suitable in applications where these properties are critical.

We applied the measures of distance to the problem of testing for a difference in diet between male and female grey seals prior to the breeding season of 2011. Using a multivariate permutation test with three different test statistics based on the ANG, MAH and CS distances between samples of QFASA diet estimates containing many zeros, we determined that there was a significant difference in the diet of the two groups of seals. Based on our simulation results, we prefer the CS over the MAH based test statistic because there were only six seals in the first sample. However, there was little difference in the three p-values.

Acknowledgements

The author would like to acknowledge summer student Dan McMullen and Chris Field for his constructive feedback on the manuscript. The author is also grateful to biologists Shelley Lang, Sara Iverson, and Don Bowen for providing and helping with the real-life seal data set. We would also like to thank the reviewers who provided important feedback which improved the quality of the manuscript. This work was supported by the Natural Sciences and Engineering Research Council of Canada.

References

[1] J. Aitchison, *On criteria for measures of compositional difference*, Mathematical Geology 24 (1992), pp. 365–379.

[2] C.A. Beck, S.J. Iverson, W.D. Bowen, and W. Blanchard, *Sex differences in grey seal diet reflect seasonal variation in foraging behavior and reproductive expenditure: evidence from quantitative fatty acid signature analysis*, Journal of Animal Ecology (2007), pp. 490–502.

[3] S.M. Budge, S.J. Iverson, W.D. Bowen, and R.G. Ackman, *Among-and within-species variation 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 (2002), pp. 886–898.

[4] J. Egozcue and V. Pawlowsky-Glahn, *Basic concepts and procedures*, in *Compositional Data Analysis: Theory and Applications*, V. Pawlowsky-Glahn and A. Buccianti, eds., John Wiley and Sons, Ltd, New York, 2011, pp. 12–28.

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

[6] M. Greenacre, *Compositional data and correspondence analysis*, in *Compositional Data Analysis: Theory and Applications*, V. Pawlowsky-Glahn and A. Buccianti, eds., John Wiley and Sons, Ltd, New York, 2011, pp. 104–113.

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

[8] S.J. Iverson, C. Field, W.D. Bowen, and W. Blanchard, *Quantitative fatty acid signature analysis: A new method of estimating predator diets*, Ecological Monographs 72 (2004), pp. 211–235.

[9] D. Jackson, *Compositional data in community ecology: The paradigm or peril of proportions?*, Ecology 78 (1997), pp. 929–940.
[10] R. Johnson and D. Wichern, *Applied Multivariate Statistical Analysis*, Pearson Education, Inc., Upper Saddle River, NJ, 2007.
[11] P.E. Kirsch, S.J. Iverson, and W.D. Bowen, *Effect of a low-fat diet on body composition and blubber fatty acids of captive juvenile harp seals (Phoca groenlandica)*, Physiological and Biochemical Zoology 73 (2000), pp. 45–59.
[12] J. Martín-Fernández, C. Barceló-Vidal, and V. Pawlowsky-Glahn, *Measures of difference for compositional data and hierarchical clustering methods*, in *Proceedings of IAMG*, 1998.
[13] J.A. Martín-Fernandez, J. Palarea-Albaladejo, and R.A. Olea, *Dealing with zeros*, in *Compositional Data Analysis: Theory and Applications*, V. Pawlowsky-Glahn and A. Buccianti, eds., John Wiley and Sons, Ltd, New York, 2011, pp. 43–58.
[14] B.H. McArdle and M.J. Anderson, *Fitting multivariate models to community data: A comment on distance-based redundancy analysis*, Ecology 82 (2001), pp. 290–297.
[15] J. Palarea-Albaladejo and J.A. Martín-Fernández, *Values below detection limit in compositional chemical data*, Journal of Analytica Chimica Acta 764 (2013), pp. 32–43.
[16] D.A.S. Rosen and D.J. Tollit, *Effects of phylogeny and prey type on fatty acid calibration coefficients in three pinniped species: implications for the QFASA dietary quantification technique*, Marine Ecology Progress Series 467 (2012), pp. 263–276.
[17] C. Stewart and C. Field, *Managing the essential zeros in quantitative fatty acid signature analysis*, Journal of Agriculture, Biological and Environmental Statistics 16 (2011), pp. 45–69.
[18] C. Stewart, S. Iverson, and C. Field, *Testing for a change in diet using fatty acid signatures*, Environmental and Ecological Statistics 21 (2014), pp. 775–792, 10.1007/s10651-014-0280-9.
[19] Y. Takane, W. Young, and J. De Leeuw, *Non-metric individual differences multidimensional scaling: Alternating least squares with optimal scaling features.*, Psycometrika 42 (1977), pp. 7–67.
[20] J. **Palarea** Albaladejo, J.A. Martín-Fernández, and J. Soto, *Dealing with distances and transformations for fuzzy c-means clustering of compositional data*, Journal of Classification 29 (2012), pp. 744–169.
[21] M. Tsagris, S. Preston, and A. Wood, *A data-based power transformation for compositional data*, in *Proceedings of the 4th International Workshop on Compositional Data Analysis*, 2011.