Compositional observations and their role in regression

Cal Spracklen (SPRCAL001@myuct.ac.za)

25th January 2018

Abstract

This dissertation examines the properties and some of the uses of compositional data. It gives a brief history of the distinction between 'normal' data and compositions, as well as the various methods of analysing compositional data. It is mainly concerned with performing regression analysis including compositions. In order to model a composition it is necessary to understand the nature of compositions and how to use standard statistical tools with them. This dissertation describes the simplex and several functions which can be performed in it, as well as introducing several useful covariate structures for compositional samples after transformation. It also introduces the transformations between the simplex and unconstrained real space. The dissertation concludes with four examples of regression analyses involving compositions.

Contents

Chapter 1

An introduction to compositional data

This chapter explains what a composition is and gives a numerical example of one. The most important properties of compositions are then stated and tools for visualising them are given. The importance of knowing which 'scale' is to be used with the analysis is then discussed. Scale refers here to the sample space of the data and the operations that could be used in this space (commonly chosen to be the simplex with the operations of perturbation and powering: see Section 2.2).

1.1 Compositions

Compositional data are data which contain only relative information. Data of this type occur when analysing parts of a whole. Compositional techniques can be useful in studying the relationships between different compositional data-sets or between compositional and non-compositional data in many different disciplines. Such disciplines could be geology, economics or chemistry for example. A composition refers to a vector of positive components which sum to some pre-specified value. This value is often 1 or 100, but can be any value as long as the components always sum to the same value. It is important to use the tools relating to compositional data when the original (full) data are not available or are constrained to a specific value. There are cases where incorrect conclusions have been drawn due to incorrect methodologies. For instance, Chayes (1962) criticizes much work in petrology which ignored the constraints on percentage data.

Compositional data are a data-set consisting of several D-part compositions, with each composition summing to any predefined value. More form-

ally, compositions are defined by Pawlowsky-Glahn *et al.* (2007, p. 5) as follows.

Definition 1. *A row vector, $X = [x_1, x_2, x_3, \ldots, , x_D]$ is defined as a D-part composition if all its components are strictly positive real numbers and they carry only relative information.*

I will use the notation $d = D - 1$ for this paper. It can be noted that in the above definition all components are taken to be strictly positive, but an alternative definition would be to require non-negative components. If any component were zero, a composition would be difficult to work with. There are many attempts at overcoming this problem, although Aitchison (2005) states that essential zeros are still highly problematic. The most common method of dealing with a zero value is to simply set the zero components to a value slightly greater than zero, and reduce the other components by an appropriate amount. I will use this convention if necessary. Aitchison (1982) introduces an important principle of compositional data analysis, and Aitchison (2003) gives two more. These three are the principles of scale invariance, subcompositional coherence and permutation invariance. A brief discussion of each of these principles will now be given.

The principle of 'scale invariance' states that compositional data carry only relative information. It tells us neither how the absolute amounts differ nor why they differ. Table 1.1 below represents the daily macronutrient intakes of an individual (subject 1) in grams, extracted from Simpson *et al.* (2003).

Table 1.1: Table of absolute consumptions (in grams) for subject 1

Day	Protein eaten	Fat eaten	Carbohydrate eaten
1	118	55	277
2	106	76	400
3	146	50	94
4	134	53	80
5	88	32	348
6	117	64	423

From Table 1.1 we can see how the amounts of each component consumed on the relevant days differ. We can compare this to the purely compositional data. Table 1.2 consists of the corresponding compositional values for each of the days.

When considering a purely compositional data-set, the only information available is of the relative sizes of components. If Table 1.2 were our only

Table 1.2: Table of compositional consumptions implied by Table 1.1

Day	Protein eaten	Fat eaten	Carbohydrates eaten
1	0.2622	0.1222	0.6156
2	0.1821	0.1306	0.6873
3	0.5034	0.1724	0.3241
4	0.5019	0.1985	0.2996
5	0.188	0.0684	0.7436
6	0.1937	0.106	0.7003

source of data, we would be unable to draw any conclusions about the absolute differences between the components. We would not be able to tell whether the change from day one to two in the proportion of carbohydrates eaten was due to an absolute increase in carbohydrates eaten or an absolute decrease in protein and fat consumed. Pawlowsky-Glahn *et al.* (2007, pp. 7, 8) explain that typically the absolute data are unavailable when compositional analysis is performed. The problem with this is that it is unknown whether the differences are due to an increase in some distinct value or a decrease in some other value. They go on to explain equivalent classes and give the following definition for compositional equivalence.

Definition 2. *Two vectors of D positive real components are compositionally equivalent if there exists a positive scalar such that $X = \lambda Y$.*

The second important principle of compositional analysis, explained by Aitchison (2003), is that of 'subcompositional coherence'. This principle states that any method used should produce consistent results between a full data-set and a subset obtained by deleting some components. For example, one scientist could be interested in the relations between protein, fat and carbohydrates as given in the table above, and another only interested in protein and fat. These two scientists should draw consistent conclusions about the interactions of protein and fat. That is, any statements made by the two scientists about any components common to the data-sets should be consistent. When analysts ignore this principle, erroneous conclusions can be drawn. This is avoided if ratios between components are analysed, or scale-invariant functions used. A simple example of ignoring this principle would be to look at the correlation coefficient in a subset of a composition. For the original full composition the correlation coefficient between fat and carbohydrates is -0.9293. The correlation coefficient between fat and carbohydrates of a subset of the data consisting of only fat and carbohydrates is necessarily -1.

These coefficients differ and thus the principle of subcompositional coherence is not observed by such correlations.

The principle of 'permutation invariance', described by Aitchison (1982), tells us that the order in which the data appear in a composition should not matter; as long as all compositions are ordered in a consistent manner, any conclusions will be independent of ordering.

1.2 A brief history of compositions

This section gives an account of how compositional analysis has evolved over the years. It introduces the different stages that compositional analysis has gone through and discusses how the different aspects of the analysis have changed. Bacon-Shone (2011) provides a short history of compositional data (which is where many of the sources in this section come from, along with a useful list of references).

1.2.1 The four phases of the history of compositional analysis

Aitchison (2005) describes the four most prominent stages of the analysis of compositional data. The initial phase of analysis failed to recognise compositions as their own data class. As a result of this failure, they were often analysed within $\mathbb{R}^d$. The regular multivariate tools were used on the data and no effort was made to compensate for the constrained nature of the data. This often lead to incorrect conclusions being drawn due to incorrect assumptions. The second stage was initiated by Chayes (1960), who drew attention to the negative bias and spurious correlation problems within compositional data. He did not provide adequate solutions to these problems. The solutions Chayes (1960) provided led to the use of distorted multivariate techniques when working with compositions. The third phase was started by Aitchison (1986) explaining that compositions provide useful information only about relative values. He suggested the use of ratios and log-ratios to analyse compositions and developed transformations for the data. These transformations were accepted and have been widely used. In the fourth stage, an understanding of the simplex, and the ability to work within this space were developed. This work includes defining operations within the simplex, so that analysts need not leave the simplex for certain problems.

1.2.2 The spurious correlation problem

Pearson (1897) is often credited with the birth of compositional data analysis. In his paper he described the problem of spurious correlation between ratios of variables. That is even if components P, Q and R are uncorrelated, P/R and Q/R will almost always have some correlation. If ignored this could lead to erroneous statements being made about the components. Pearson (1897) studied the use of adjusting the correlations in order to remedy this problem. Tanner (1949) suggested that log transformed data may avoid the spurious correlation problems. Chayes (1960) then made a direct connection between Pearson's work and compositional analysis. He pointed out that some correlations between components must necessarily be negative due to the unit-sum constraint, but proposed no solution for this problem.

1.2.3 The birth of the log-ratio transformations

McAlister (1897) made a very important step towards the modern view of compositional data by his use of log-normal distributions to model positively constrained real data. The log-normal distribution was more thoroughly explored by Aitchison and Brown (1969). Their text addressed the positive nature of the compositional data, but not the unit-sum constraint. Bacon-Shone (2011) reports that both Obenchain (in personal communication with Johnson and Kotz) and Kotz *et al.* (2000) discussed the use of a logit transformation to help model compositional data. This is a development of the work of Cox and Snell (1989) to model the probabilities of a binary outcome with the logit transform. Aitchison and Shen (1980) were the first to publicly introduce the logistic-normal distribution for compositional data. They described the distribution in terms of log-ratios relative to the final component. Aitchison (1986) further developed the ideas of the logistic-normal distribution. He showed that the covariance structure of a composition could be completely defined by $D(D-1)/2$ log-ratio variances, where D refers to the size of the composition.

1.2.4 The transformations

Aitchison (1986, p. 113) was the first to develop the additive log-ratio transformation for compositional data analysis. This transformation takes the log-ratios of all the components of a composition over one common component. There was some doubt about this transformation due to the arbitrary choice of a divisor component necessary for it. There were also thoughts that different divisors could lead to different conclusions. Aitchison (1986) and

Aitchison *et al.* (2000) show that this does not cause any practical problems. That is, the component chosen as a divisor for the transformation will not affect the results of an analysis. Aitchison (1986, p. 79) also developed the centred log-ratio transformation. This transformation avoids the problem of choosing any one component as the divisor. It does this by instead taking the log-ratio of each component and the geometric mean of the composition. Egozcue, Pawlowsky-Glahn, Mateu-Figueras, and Barcelo-Vidal (2003) developed the isometric log-ratio, which circumvents the issue of choosing any one divisor by instead basing the transformation on an orthonormal basis. This transformation was developed in order to preserve consistency between the simplex and $\mathbb{R}$ with respect to certain metrics.

1.2.5 Operations within the simplex

Aitchison (1986, p. 42) developed several operations for working with compositions in the simplex. The operations are analogues of the familiar addition and multiplication functions and are referred to as perturbation and powering respectively. The function analogous to addition was extended by Aitchison and Ng (2005). These functions were later used to define the simplex as its own vector space. They can also be used to define some geometry on the simplex.

1.2.6 Some areas of study

Compositional data were initially studied primarily by statisticians who had interest in how the constrained data reacted when under analysis. The field of petrology was one of the first to benefit from a thorough investigation as to how compositions are analysed. This is due to the geochemical compositions of rocks being of great importance in petrology. Aitchison (2005) suggests that there are still pockets of resistance in petrology towards a correct methodology of an analysis. Sedimentology is another field which has benefited from compositional analysis. Sediment specimens are usually separated into mutually exclusive and exhaustive components which can then be analysed. Aitchison (2003) suggests a range of fields in which the analysis of compositional data has proved useful. Aitchison (1986, pp. 285–291) also suggested the use of compositions to analyse mixture distributions. More recently work has been done in the medical field by Dumuid *et al.* (2017) and Lea and Leite (2016).

1.3 Visual Representations

This section introduces the ternary diagram, a simple tool used to represent compositions containing three components and the log-ratio scatter-plots. Compositions can also be represented by the more complex compositional biplot which will also be examined. More information on compositional biplots and biplots in general can be found in Pawlowsky-Glahn *et al.* (2015) and Aitchison and Greenacre (2002).

1.3.1 Ternary diagram and log-ratio scatter-plots

A three-part composition can be conveniently represented by a ternary diagram. This diagram is a triangle where each vertex represents a component of the composition. The compositions are plotted as points in this diagram in such a manner that the closer a point is to a vertex of the triangle, the higher the value of the corresponding component. For a four-part composition the ternary diagram could be extended into three-dimensional space. Unfortunately, due to the inability to perceive four-dimensional or higher spaces, we cannot extend the diagram any further. If we wished to represent compositions of higher order, then Aitchison (1986, p. 9) suggests that we can use multiple ternary diagrams to represent the subcompositions. The data given previously for the consumption of protein, fat and carbohydrates can be plotted into a ternary diagram, as seen in Figure 1.1.

Aitchison (2005) also suggests the use of log-ratio scatter-plots to help visualise the data. Van den Boogaart and Tolosana-Delgado (2013, pp. 27, 28) explain that these scatter-plots are just plots of the log-ratio of two components plotted against the log-ratio of another two components. (But they state that the denominator for the two plotted log-ratios may be the same.)

1.3.2 Compositional biplot

Pawlowsky-Glahn *et al.* (2015, pp. 70–76) describe how to construct and interpret a biplot of a compositional data-set. The purpose of a biplot is to represent the rows and columns of a matrix via a rank-2 approximation. Here we shall be working with n observations of a D-part composition. The corresponding $n \times D$ matrix of compositions, is referred to as $\mathbf{X}$. We construct a matrix $\mathbf{Z}$ by applying the following operation to each of the rows of $\mathbf{X}$:

$$\mathbf{Z}_i = \log(\frac{\mathbf{X}_i}{g(\mathbf{X}_i)}).$$

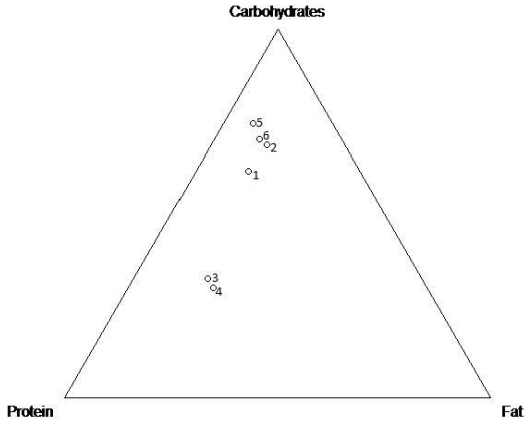


Figure 1.1: Ternary diagram for consumption of nutrients over 6 days

(In this equation $\mathbf{X}_i$ and $\mathbf{Z}_i$ refer to the rows of $\mathbf{X}$ and $\mathbf{Z}$ respectively and $g(\mathbf{X}_i)$ is the geometric mean of the ith row of $\mathbf{X}$.) The transformation applied to the rows $\mathbf{X}_i$ is referred to as the centred log-ratio (clr) transformation and is discussed in more detail later. We note here that the clr transformation preserves the rank of the data matrix, $\mathbf{X}$.

We now find the best rank-2 approximation $\mathbf{Z}^*$ to $\mathbf{Z}$ in a least squares sense. This approximation is provided by a singular value decomposition (SVD). The SVD of $\mathbf{Z}$ is written as $\mathbf{UKV}'$. $\mathbf{U}$ and $\mathbf{V}$ are the matrices of the eigenvectors of $\mathbf{ZZ}'$ and $\mathbf{Z}'\mathbf{Z}$ respectively. $\mathbf{K}$ is a diagonal matrix. The diagonal of $\mathbf{K}$ consists of the ordered square roots of the positive eigenvalues $\lambda_1, \lambda_2, \ldots, \lambda_r$ of either $\mathbf{ZZ}'$ or $\mathbf{Z}'\mathbf{Z}$ (The positive eigenvalues for these two matrices are the same). The number of positive eigenvalues, r, is $\leq \min\{d, n\}$ where d is equal to $D - 1$ and n is the number of observations. In order to construct this diagonal matrix we must first write $k_i = \lambda_i^{1/2}$ and then rearrange the k_is in descending order of magnitude (with the largest value denoted by k_1). The matrix $\mathbf{K}$ represents the diagonal matrix of k.

In order to find $\mathbf{Z}^*$ it is necessary to reduce the dimensionality of $\mathbf{Z}$. We do this by substituting 0 for k_{ii} where $i > 2$. (If we wanted a reduced matrix of different dimensionality we could substitute an appropriate number for 2 in the preceding equation. For example a rank-3 approximation could be obtained by substituting 0 for all k_{ii} where $i > 3$.)

The rank-2 approximation for $\mathbf{Z}$ follows:

$$\mathbf{Z}^{*\mathbf{2}} = \begin{pmatrix} u_{11} & u_{21} \\ u_{12} & u_{22} \\ \vdots & \vdots \\ u_{1n} & u_{2n} \end{pmatrix} \begin{pmatrix} k_1 & 0 \\ 0 & k_2 \end{pmatrix} \begin{pmatrix} v_{11} & v_{21} & \dots & v_{D1} \\ v_{12} & v_{22} & \dots & v_{D2} \end{pmatrix}.$$

To construct a biplot $\mathbf{Z}^{*\mathbf{2}}$ is rearranged into the product of two matrices. The biplot depends on a new value $p \in [0, 1]$. Aitchison and Greenacre (2002) refer to the case $p = 1$ as the covariance biplot, and the case $p = 0$ as the form biplot. In the covariance biplot the length of any vector is proportional to the standard deviation of the corresponding clr transformed component. In the form biplot the distance between any compositions in the diagram is related to the actual distance between those compositions. Below the matrix $\mathbf{Z}^{*\mathbf{2}}$ is represented as a product of two matrices:

$$\begin{aligned} \hat{\mathbf{Z}^{*\mathbf{2}}} &= \begin{pmatrix} \sqrt{n}k_1^{1-p}u_{11} & \sqrt{n}k_2^{1-p}u_{21} \\ \sqrt{n}k_1^{1-p}u_{12} & \sqrt{n}k_2^{1-p}u_{22} \\ \vdots & \vdots \\ \sqrt{n}k_1^{1-p}u_{1n} & \sqrt{n}k_2^{1-p}u_{2n} \end{pmatrix} \begin{pmatrix} \frac{k_1^p v_{11}}{\sqrt{n}} & \frac{k_1^p v_{21}}{\sqrt{n}} & \dots & \frac{k_1^p v_{D1}}{\sqrt{n}} \\ \frac{k_2^p v_{12}}{\sqrt{n}} & \frac{k_2^p v_{22}}{\sqrt{n}} & \dots & \frac{k_2^p v_{D2}}{\sqrt{n}} \end{pmatrix} \\ &= \begin{pmatrix} a_1 \\ a_2 \\ \vdots \\ a_n \end{pmatrix} \begin{pmatrix} b_1 & b_2 & \dots & b_D \end{pmatrix} \end{aligned}$$

The biplot is then constructed by plotting the vectors b_i, and points a_i on a plane. This biplot contains a large amount of information. The centre of the b_is is the origin, termed O, and represents the geometric mean of the sample.

As an illustration we now construct a biplot for the data found in Table 1.2. The first step is to transform the matrix of observations via the centred log-ratio transformation. The resulting matrix,

$$\mathbf{Z} = \begin{pmatrix} -0.0300 & -0.793 & 0.823 \\ -0.332 & -0.665 & 0.996 \\ 0.504 & -0.568 & 0.064 \\ 0.481 & -0.446 & -0.035 \\ -0.121 & -1.133 & 1.254 \\ -0.227 & -0.831 & 1.058 \end{pmatrix}$$

can then be decomposed into a SVD. The decomposition can be written as:

$$\begin{pmatrix} -0.412 & 0.085 & -0.249 \\ -0.437 & -0.295 & 0.205 \\ -0.137 & 0.687 & -0.256 \\ -0.083 & 0.642 & 0.550 \\ -0.612 & 0.028 & -0.452 \\ -0.481 & -0.145 & 0.569 \end{pmatrix} \begin{pmatrix} 2.770 & 0 & 0 \\ 0 & 0.960 & 0 \\ 0 & 0 & 2 \times 10^{-16} \end{pmatrix} \begin{pmatrix} 0.084 & 0.812 & -0.577 \\ 0.661 & -0.479 & -0.577 \\ -0.746 & -0.333 & -0.577 \end{pmatrix}.$$

We now perform a reduction of the SVD resulting in:

$$\begin{pmatrix} -0.412 & 0.085 \\ -0.437 & -0.295 \\ -0.137 & 0.687 \\ -0.083 & 0.642 \\ -0.612 & 0.028 \\ -0.490 & -0.145 \end{pmatrix} \begin{pmatrix} 2.770 & 0 \\ 0 & 0.960 \end{pmatrix} \begin{pmatrix} 0.084 & 0.812 & -0.577 \\ 0.661 & -0.479 & -0.577 \end{pmatrix}.$$

Finally we can convert these matrices into the vectors, b_i, and points, a_i, for the biplot using the equations given above. The covariance biplot can be seen in Figure 1.2.

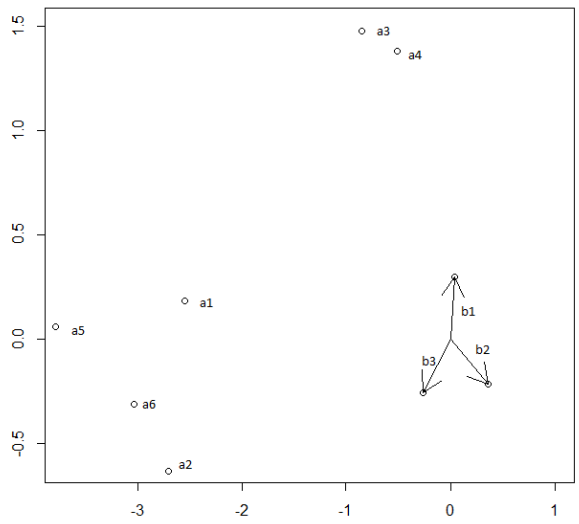


Figure 1.2: Covariance biplot, with each point (a_i) representing a composition and each vector (b_i) a component of the compositions.

1.3.3 Interpreting the biplot

In order to ease the explanation I will define several terms associated with a biplot, as done by Pawlowsky-Glahn *et al.* (2015, p. 77). The centre of the biplot, that is the start of all the vectors, is termed the origin. The points described by the a_is are referred to as the case markers, and the b_is as vertices. The vector joining the origin to a vertex is termed a ray, and the join of any two vertices is called a link.

Links and rays can give an idea of the variability of a log-ratio in the compositional data-set when seen in the covariance biplot. The squared length of the origin to a vertex is approximately equal to the variance of the log of that component over the geometric mean, that is $|Ob_i|^2 \approx \text{var}(\log(\frac{x_i}{g(x)}))$. The distance between any two vertices squared is approximately equal to the variance of the log-ratio of the components associated with those vertices, $\text{var}(\log(\frac{x_i}{x_j}))$.

The distances between the points in a form biplot represent the compositional distance between the corresponding compositions. The compositional distance will be further defined in a later chapter.

The angles between links in the covariance biplot can be used to estimate the correlation between log-ratios. The cosine of the angle between the links is an approximation of the correlation between the log-ratio of the corresponding components for the vertices. If M represents the angle of the intercept of the links between b_j, b_k and b_i, b_l then $\cos(M) \approx \text{corr}(\log(\frac{x_j}{x_k}), \log(\frac{x_i}{x_l}))$. These are the basic properties of a compositional biplot. Pawlowsky-Glahn *et al.* (2015) and Aitchison and Greenacre (2002) provide more information on the interpretation of a biplot.

1.4 Scales

A fundamental property of any set of variables in a data-set is its 'scale'. The term scale here refers to both the set of all possible outcomes of values and meaningful mathematical operations. Each scale will usually have statistical models associated with it. Compositional data will lie on some multivariate scale, as all of the components are only meaningful in relation to each other. Compositional data have been treated in several different ways and have been worked with on different scales. Van den Boogaart and Tolosana-Delgado (2013, pp. 34, 35) suggest that it is necessary to select the most meaningful scale for the problem. They state a preference for the Aitchison compositional scale and suggest that this be the default choice.

Van den Boogaart and Tolosana-Delgado (2013, pp. 31–34) describe five main scales which can be used for analysing compositions.

1. Classical multivariate vectorial data are the scale representing the Cartesian product of D real scales, with values in $\mathbb{R}^D$. The mathematical operations are defined as one would expect in $\mathbb{R}^D$. The problem with this scale is that the values are not strictly compositional, nor are they strictly positive. Most multivariate statistical tools work with this scale and it is often used in compositional analysis after a transformation has been used on the data.

2. Positive data with absolute geometry are a scale with values in $\mathbb{R}_+^D$, the positive orthant of $\mathbb{R}^D$. The usual statistical tools when working with positive data can be used when working in this scale. There may be some difficulty in interpreting the results of an analysis within this scale, as the results may be nonsensical. Van den Boogaart and Tolosana-Delgado (2013, p. 31) point out that an outcome with a negative prediction for a proportion could occur from using this scale.

3. Positive data with relative geometry are a ratio scale used when comparing relative observations. It is commonly used after log transformation of the underlying data. The basic idea when using this scale is to analyse each component individually using a relative scale. The products, divisions and distances of the composition would be calculated using the log-transformed components.

4. Compositional data with absolute geometry involve treating compositional data as if they were subsets of a multivariate real data-set. This is the most commonly used of the possible scales. Working in this scale involves transforming the composition into $\mathbb{R}^d$ space and using regular statistical methods.

5. Compositional data with Aitchison geometry attempt to honour the basic principles of compositional analysis. They obey the principles of scale invariance, permutation invariance and subcompositional coherence. They also use the ratios of compositions. The reference distribution is the additive logistic normal distribution (A reference distribution is one which is used to transform the properties of one sample space into another). The vector space associated with the Aitchison geometry (the simplex S^d) is isometrically equivalent to $\mathbb{R}^d$.

Chapter 2

Geometry of the simplex

This chapter starts by introducing the simplex, which is the space on which a composition is defined. It will then define some operations on the simplex, two of which are analogous to addition and scalar multiplication. Examples of these operations will be provided. It is worth a reminder that in this dissertation d is defined as $D - 1$.

2.1 Form of the simplex

Aitchison (1986, p. 27) provides two definitions of the simplex in which a composition is found. The first definition is:

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

This form could be used when we are more concerned with specifying density functions or other problems related to the dimensionality of the composition. This definition leads us to a solid object in $\mathbb{R}^d$. The second, more common and probably more useful, definition is:

$$S^d = \{(x_1, \ldots, x_D) : x_1 > 0, \ldots, x_D > 0; x_1 + \ldots + x_D = L\}.$$

This is a more symmetric definition as it treats all components equally. This second definition describes a hyperplane in $\mathbb{R}^D$.

2.2 Operations within the simplex

Some operations are defined in order to describe a composition without transforming the composition into unconstrained d-dimensional space, $\mathbb{R}^d$. These

operations describe the geometry in the simplex and allow a discussion of operations within the simplex.

In order to illustrate these operations I introduce the following data from Simpson *et al.* (2003). The data are provided in Table 2.1, which represents the amount of macronutrients consumed in grams by an individual (subject 2).

Table 2.1: Table of absolute consumptions for subject 2.

Day	Protein eaten	Fat eaten	Carbohydrates eaten
1	96	62	271
2	114	92	402
3	130	49	73
4	127	57	73
5	92	73	415
6	111	101	443

Pawlowsky-Glahn *et al.* (2007, p. 6) define the **closure** operation $C()$, which is used to rescale a vector z so that it lies within the simplex. When a D-part composition in the simplex is required to sum to f, then C is defined by

$$C(z) = \left(\frac{fz_1}{z_1 + z_2 + \ldots + z_D}, \frac{fz_2}{z_1 + z_2 + \ldots + z_D}, \ldots, \frac{fz_D}{z_1 + z_2 + \ldots + z_D} \right).$$

For instance the data in Table 2.1 becomes the data in Table 2.2 when closure is applied with a unit sum constraint, that is closure is applied with $f = 1$.

Table 2.2: Table of compositional consumptions implied by Table 2.1.

Label	Day	Protein eaten	Fat eaten	Carbohydrates eaten
$D1$	1	0.2238	0.1445	0.6317
$D2$	2	0.1874	0.1514	0.6612
$D3$	3	0.5159	0.1944	0.2897
$D4$	4	0.4942	0.2218	0.2840
$D5$	5	0.1586	0.1259	0.7155
$D6$	6	0.1695	0.1542	0.6763

Van den Boogaart and Tolosana-Delgado (2013, pp. 37–40) define several operations necessary to explore the data within the composition. They

start by defining **perturbation** as a compositional sum within this simplex. Perturbation is, for two D-part compositions X and Y: $Z = X \oplus Y = C(x_1y_1, \ldots, x_Dy_D)$. The identity element for such an operation is easily shown to be the composition $(1/D, \ldots, 1/D)$ and the inverse of X under perturbation is $\ominus X = C(1/x_1, \ldots, 1/x_D)$. Perturbation is analogous to $+$ in $\mathbb{R}^d$. This equivalence is due to the same conclusions being reached after certain transformations to the data that are discussed in detail in Chapter 3. These transformations are namely the additive log-ratio transformation, the centred log-ratio transformation and the isometric log-ratio transformation. For example, applying $\oplus$ to two compositions and transforming the result will give the same answer as transforming two compositions and summing them. To numerically illustrate perturbation we can calculate the perturbation of the compositions corresponding to days one and two for Subject 2:

$$\begin{aligned} D1 \oplus D2 &= C(0.2238 \times 0.1874, 0.1445 \times 0.1514, 0.6317 \times 0.6612) \\ &= (0.08710, 0.04544, 0.86746). \end{aligned}$$

The operation **powering** denoted by $\odot$ is defined as follows:

$$\lambda \odot X = C(x_1^\lambda, \ldots, x_D^\lambda)$$

for some scalar λ and composition X with components $x_1, \ldots, x_D$. Powering is analogous to multiplication in real space.

If we power the composition found in day three of Table 2.2 by a factor of four the following result is observed:

$$4 \odot D3 = C(0.5159^4, 0.1944^4, 0.2897^4) = (0.8932, 0.0180, 0.0888).$$

The simplex, $(S^d, \oplus, \odot)$, with perturbation $(\oplus)$ and powering $(\odot)$, is a vector space. Pawlowsky-Glahn *et al.* (2015, pp. 24–26) show this, in particular they prove the following properties. $(S^d, \oplus)$ is a commutative group and powering satisfies the properties of an inner product.

Pawlowsky-Glahn *et al.* (2015, pp. 26, 27) provide on $(S^d, \oplus, \odot)$ a compositional inner product and hence a norm and metric for compositions. The Aitchison inner product, defined as $< X, Y >_A = \frac{1}{2D} \sum_{i=1}^{D} \sum_{j=1}^{D} \log(\frac{x_i}{x_j}) \log(\frac{y_i}{y_j})$, has several alternative but equivalent forms but this is the most widely used. The norm of a vector, referred to as the Aitchison norm, is $||X||_A = \sqrt{< X, X >_A}$. Finally the Aitchison distance is $d_A(X, Y) = ||X \ominus Y||_A$. This

can be simplified in the following manner:

$$\sqrt{\frac{1}{2D}\sum_{i=1}^{D}\sum_{j=1}^{D}\left(\log(\frac{x_i \ominus y_i}{x_j \ominus y_j})\right)^2} = \sqrt{\frac{1}{2D}\sum_{i=1}^{D}\sum_{j=1}^{D}\left(\log(\frac{x_i y_j}{x_j y_i})\right)^2}$$
$$= \sqrt{\frac{1}{2D}\sum_{i=1}^{D}\sum_{j=1}^{D}\left(\log(\frac{x_i}{x_j}) - \log(\frac{y_i}{y_j})\right)^2}.$$

(Note the use of the $\ominus$ in the above formula is due to the geometry of the simplex and corresponds to the use of a minus sign in real space.)

The angle between compositions is defined by Van den Boogaart and Tolosana-Delgado (2013, p. 40) as $\alpha(X,Y)_A = \arccos(\frac{<X,Y>_A}{||X||_A||Y||_A})$ with range $0°$ to $180°$.

In order to demonstrate the above operations a numerical example is provided. For this example X represents the composition for day four and Y the composition for day five, with data drawn from Table 2.2.

Aitchison inner product	$\frac{1}{6}\sum_{i=1}^{3}\sum_{j=1}^{3}\left(\log(\frac{x_i}{x_j})\log(\frac{y_i}{y_j})\right)$ $=-0.07337042$
Aitchison distance	$\sqrt{\frac{1}{6}\sum_{i=1}^{3}\sum_{j=1}^{3}\left(\log(\frac{x_i}{x_j})\log(\frac{y_i}{y_j})\right)^2}$ $=1.504671$
Angle between compositions	$\arccos\left(\frac{-0.07337042}{||x||_A||y||_A}\right)$ $= 1.665707$

With the above operations Pawlowsky-Glahn *et al.* (2015, pp. 28–30) explain what compositional lines and compositional circles are. A compositional line is defined as a line lying on the simplex. This line takes the form of $Y = X_0 \oplus (\alpha \odot X)$, where X_0 is the starting composition, X is a directional composition and α is a real scalar. A compositional circle is a circle defined on the simplex. This circle takes the form $r = d_A(X, X_0)$, where $r \in \mathbb{R}^+$ is the radius of the compositional circle and X and X_0 are compositions defined on the simplex.

Chapter 3

Transformations from the simplex to unconstrained real space

This chapter examines the transformations between S^d and $\mathbb{R}^d$. Aitchison (2003) gives two transformations from the simplex to unconstrained real space, namely the additive log-ratio transformation and the centred log-ratio transformation. Van den Boogaart and Tolosana-Delgado (2013, p. 42) introduce a third, the isometric log-ratio transformation. It is important to note that, under all of these transformations, the operations of powering and perturbation translate to multiplying and addition respectively.

For this chapter, X and Y will represent D-part compositions, and z a vector in $\mathbb{R}^d$. For the numerical transformations I will be using the compositions given in Table 1.2 for days five and six. These will be termed $P = (0.188, 0.0684, 0.7436)$ and $Q = (0.1937, 0.106, 0.7003)$.

3.1 Additive log-ratio transformation

The first transformation described is the additive log-ratio transformation (alr). This transformation stems from the idea that we are concerned only with the relative proportions of the compositions, not their actual sizes. The transformation is defined by:

$$\text{alr}(X) = [\log(x_1/x_D), \log(x_2/x_D), \ldots, \log(x_d/x_D)],$$

with an inverse operation of $\text{alr}^{-1}(z) = C[e^{z_1}, e^{z_2}, \ldots, e^{z_d}, 1]$. The alr transformation transforms the data from the simplex, S^d, into the real space of $\mathbb{R}^d$. No isometry between S^d and $\mathbb{R}^d$ is provided by this transformation. (An

isometry is a distance-preserving mapping between two metric spaces.) This is due to the Aitchison inner product not being preserved under the alr transformation, that is $< X, Y >_A \neq \text{alr}(X)\text{alr}(Y)^T$. The transformation divides each of the components by a common component. This component can be any of those in the composition. (The log of this ratio is then calculated and the outcome of this calculation will be the transformed data.) The final element of the composition is often chosen as the denominator. This operation is of a one-to-one nature and thus any inferences can be transformed back to the simplex. The choice of the divisor is irrelevant for an analysis of the composition; Aitchison (1986, pp. 78) states that any of the components could be used as the denominator. The main drawback of this transformation is its lack of distance preservation.

Example 3.1. Using the alr transformation we can find $\text{alr}(P)$ and $\text{alr}(Q)$ respectively.

$$\begin{aligned} \text{alr}(P) &= \left(\log(\frac{0.188}{0.7436}), \log(\frac{0.0684}{0.7436})\right) \\ &= (-1.375061, -2.386130) \\ \text{alr}(Q) &= \left(\log(\frac{0.1937}{0.7003}), \log(\frac{0.106}{0.7003})\right) \\ &= (-1.285198, -1.888070). \end{aligned}$$

□

3.2 Centred log-ratio transformation

The centred log-ratio transformation (clr), described by Aitchison (2003), transforms the data from the simplex to a hyperplane of $\mathbb{R}^D$. (This hyperplane can be defined as $U^D = [u_1, \ldots, u_D] : u_1 + \ldots + u_D = 0$.) This transformation removes the problem of choosing a component as the divisor and instead divides each component by the geometric mean of the composition and then calculates the log of the ratio. The transformation takes the form

$$\text{clr}(X) = [\log(x_1/g(x)), \log(x_2/g(x)), \ldots, \log(x_D)/g(x)].$$

Here $g(x)$ is the geometric mean of the composition being transformed, $g(x) = (\prod_{i=1}^{D} x_i)^{1/D}$. Pawlowsky-Glahn *et al.* (2015, p. 32) explain that when analysing random samples the covariance matrix of the data thus transformed is singular. Another problem with the clr transformation is that it is not necessarily subcompositionally coherent, which could be caused by a partition of a composition having a different geometric mean to the full composition.

Example 3.2. Using the above transformation we define $\text{clr}(P)$ and $\text{clr}(Q)$ as the clr transformation for P and Q respectively.

$$\begin{aligned}
\text{clr}(P) &= \left(\log(\frac{0.188}{0.2122517}), \log(\frac{0.0684}{0.2122517}), \log(\frac{0.7436}{0.2122517})\right) \\
&= (-0.1213309, -1.1324000, 1.2537304) \\
\text{clr}(Q) &= \left(\log(\frac{0.1937}{0.2431681}), \log(\frac{0.106}{0.2431681}), \log \frac{0.7003}{0.2431681}\right) \\
&= (-0.2274424, -0.8303139, 1.0577558)
\end{aligned}$$

□

3.3 Isometric log-ratio transformation

The isometric log-ratio transformation (ilr) is an isometric linear mapping between the simplex and $\mathbb{R}^d$. The transformation is given by Van den Boogaart and Tolosana-Delgado (2013, p. 43). (An isometric linear map is one which preserves distances and angles between points.) Pawlowsky-Glahn *et al.* (2015, p. 34) explain that the isometric log-ratio transformation was developed to preserve the Aitchison inner product, norm and distance. The ilr transformation is a fairly involved process with several steps.

The transformation is performed by first finding an orthonormal basis for S^d and transforming it into a 'contrast matrix'. The contrast matrix, denoted by Ψ, is then used to define the ilr transformation, as follows:

$$\text{ilr}(X) = \text{clr}(X)\Psi^T.$$

The covariance matrix of a sample transformed by this transformation is of full rank. The inverse operation ilr^{-1} is given by $\text{ilr}^{-1}(z) = C(\exp(z\Psi))$.

3.3.1 Finding a contrast matrix

To obtain a contrast matrix, Ψ, one needs to first find an orthonormal basis and transform this using the centred log-ratio transformation. We let V denote an orthonormal basis on the simplex and define Ψ, the contrast matrix, as a $d \times D$ matrix with ith row $\psi_i = \text{clr}(v_i)$. v_i represents the ith vector of the orthonormal basis. An orthonormal set is an orthogonal set of unit vectors. The contrast matrix formed on the simplex has the following properties: $\Psi\Psi^T = I_d$ and $\Psi^T\Psi = I_D - D^{-1}\mathbf{1_{DxD}}$. The rows of Ψ all sum to zero, that is $\Psi\mathbf{1} = 0$, where $\mathbf{1}$ is a column vector of ones.

Pawlowsky-Glahn *et al.* (2015, p. 33) explain that in order to find an appropriate basis for S^d we must first find an appropriate 'generating system'. They use the generating system $W = \{w_1, \ldots, w_D\}$, where $w_i = C[1, 1, \ldots, e, \ldots, 1]$, with $i = 1, \ldots, D$, and the e appears in the i th position. Using this generating system we can rewrite any composition $X \in S^d$ as $X = \bigoplus_{i=1}^{D} (\log(x_i) \odot w_i)$. This representation of x also implies the following result: $X = \bigoplus_{i=1}^{D} \left(\log(\frac{x_i}{g(x)}) \odot w_i\right)$, where $g(x)$ is the geometric mean of the composition.

Example 3.3. We demonstrate here the use of the generating system $W = \{C(e, 1, 1), C(1, e, 1), C(1, 1, e)\}$ to represent the composition $P = (0.1880, 0.0684, 0.7436)$. In this example the closure function present in $\odot$ is not applied as it is made redundant by being applied later.

$$\begin{aligned} & (\log(0.1880) \odot [e^1, 1, 1] \oplus \log(0.0684) \odot [1, e^1, 1] \oplus \log(0.7436) \odot [1, 1, e^1]) \\ = \; & (e^{\log(0.1880)}, 1, 1) \oplus (1, e^{\log(0.0684)}, 1) \oplus (1, 1, e^{\log(0.7436)}) \\ = \; & (0.1880, 0.0684, 0.7436) \\ = \; & P. \end{aligned}$$

□

Pawlowsky-Glahn *et al.* (2015, p. 36) state that, if any one of the elements of a generating system is omitted, a basis is obtained; e.g., the set $\{w_1, \ldots, w_d\}$ forms a basis. This basis is however not necessarily orthonormal, but the Gram–Schmidt procedure can be used to obtain an orthonormal one. (It is important to remember that, when we use the Gram–Schmidt procedure here, we must use the operations of perturbation and powering as well as the Aitchison inner product ($< X, Y >_A$).)

Example 3.4. The generating system for a 3-part composition is

$$\{C(e^1, 1, 1), C(1, e^1, 1), C(1, 1, e^1)\}.$$

A basis for S^2 can be created by removing one of the components. This yields $\{C(e^1, 1, 1), C(1, e^1, 1)\}$. The scalar product between these vectors is $2e^1 + 1$ which is not 0 and thus the basis is not orthonormal. □

Pawlowsky-Glahn *et al.* (2015, p. 36) define the contrast matrix associated with the orthonormal basis $[e_1, \ldots, e_d]$, which is used in the transformation of the composition. This matrix is defined as $\Psi = (\psi_i)$, where $\psi_i = \text{clr}(e_i)$.

Once Ψ has been found it is easy to perform the ilr transformation.

Example 3.5. We can construct an orthonormal basis for P by applying the Gram–Schmidt procedure to the basis after applying the closure operation. Here the division by a scalar and the symbol $\ominus$ represent the inverses of powering and perturbation respectively, where applicable.

$$\begin{aligned} w_1 &= C(e^1, 1, 1) = (0.5761, 0.2119, 0.2119) \\ w_2 &= C(1, e^1, 1) \ominus \frac{C(1, e^1, 1).C(e^1, 1, 1)}{||C(e^1, 1, 1)||^2} C(e^1, 1, 1) \\ &= C(1, e^1, 1) \ominus 0.7920459 C(e^1, 1, 1) \\ &= (0.3072, 0.5065, 0.1863) \end{aligned}$$

The set containing w_1 and w_2 is orthogonal, but not orthonormal. The orthonormal set consists of v_1 and v_2, which are defined by

$$\begin{aligned} v_1 &= \frac{1}{||w_1||} w_1 = \frac{1}{\sqrt{2 + 2e^1}} (C(e^1, 1, 1)) \\ &= (0.6299, 0.1851, 0.1851) \\ v_2 &= \frac{1}{\sqrt{5.8901}} (-1.1530, 1.9262, 0.2080) \\ &= (0.2840, 0.5760, 0.1400). \end{aligned}$$

The Ψ matrix consists of the clr transforms of v_1 and v_2 for each of its rows. Thus

$$\Psi = \begin{pmatrix} 0.8165 & -0.4082 & -0.4082 \\ 0 & 0.7071 & -0.7071 \end{pmatrix}.$$

We can now perform the ilr transformation by multiplying the clr transformed compositions by the transpose of Ψ, thus

$$\text{ilr}(P) = (-0.1485992, -1.687249) \text{ and } \text{ilr}(Q) = (-0.2785587, -1.335067).$$

□

3.3.2 Alternative ways of finding the contrast matrix

This section will briefly discuss three alternative ways of finding the contrast matrix needed for the ilr transformation. Egozcue *et al.* (2003) confirm that the use of a different basis or contrast matrix in the ilr transformation will yield the same properties as any ilr transformation with a differing contrast matrix. They also confirm that any analysis on the same composition with differing basis will yield consistent results.

Van den Boogaart and Tolosana-Delgado (2013, p. 43) give an alternative method to find an orthogonal basis, which is used in the ilr transformation;

they suggest attempting to choose a basis that is related in some way to the purpose of the analysis, but they admit that this is not always possible. In the event that a basis with a connection to the data cannot be found then they recommend using the method described in the previous section in order to find a basis.

Egozcue *et al.* (2003) construct a basis which can be used to find the contrast matrix by combining $D-1$ mutually orthogonal balance elements. A balance element is defined as $h_i = C(e^{\sqrt{\frac{1}{i(i+1)}}}, \ldots, e^{\sqrt{\frac{1}{i(i+1)}}}, e^{-\sqrt{\frac{1}{i(i+1)}}}, e^0, \ldots, e^0)$, where there are i terms of $e^{\frac{1}{i(i+1)}}$.

Pawlowsky-Glahn *et al.* (2015, pp. 38–41) describe a third method of finding Ψ, the contrast matrix. They recommend creating a sign table. The first row of this table is created by splitting the composition into two groups; each of these groups is then split into two more groups. This process continues until each group has a single part. A 1 is used to indicate an inclusion in the first group, -1 indicates an inclusion in the second group and 0 indicates no inclusion.

For each row we denote r by the number of $+$ signs and s by the number of $-$ signs. The contrast matrix Ψ is then constructed by the following formulae: $\psi_{ij} = \frac{1}{r}\sqrt{\frac{rs}{r+s}}$ when the corresponding value in the sign table is positive and $\psi_{ij} = \frac{1}{s}\sqrt{\frac{rs}{r+s}}$ when the corresponding value in the sign table is negative. If the corresponding sign table value is 0 then the value in the contrast matrix is also 0.

Example 3.6. Consider a 6-part composition with components $x_1, \ldots, x_6$. The first level will be constructed by splitting the composition in half. The first group is assigned $+1$ and the latter group -1. For the next level we consider only the $+1$ group and then split this, we assign 0 to the -1 group who are no longer included in either group. We continue until there is only one $+1$ in the group and then look at the -1 groupings. The sign table will look as follows:

x_1	x_2	x_3	x_4	x_5	x_6	r	s
$+1$	$+1$	$+1$	-1	-1	-1	3	3
$+1$	$+1$	-1	0	0	0	2	1
$+1$	-1	0	0	0	0	1	1
0	0	0	$+1$	$+1$	-1	2	1
0	0	0	$+1$	-1	0	1	1

After the sign table has been constructed, the Ψ matrix can be found by applying the formulae given in the main text for each of the i,j elements of the sign table where a positive or negative value appears and a 0 when the

element is 0. The resulting matrix is below. The values for r and s in the above table correspond to those of the ith row.

$$\Psi = \begin{pmatrix} 0.5 & 0.5 & 0.5 & -0.5 & -0.5 & -0.5 \\ \frac{\sqrt{2}}{2\sqrt{3}} & \frac{\sqrt{2}}{2\sqrt{3}} & -\frac{\sqrt{2}}{\sqrt{3}} & 0 & 0 & 0 \\ \frac{1}{\sqrt{2}} & -\frac{1}{\sqrt{2}} & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & \frac{\sqrt{2}}{2\sqrt{3}} & \frac{\sqrt{2}}{2\sqrt{3}} & -\frac{\sqrt{2}}{\sqrt{3}} \\ 0 & 0 & 0 & \frac{1}{\sqrt{2}} & -\frac{1}{\sqrt{2}} & 0 \end{pmatrix}$$

If we consider the composition $L = (0.3, 0.1, 0.2, 0.4, 0.05, 0.05)$ to be transformed via the ilr transformation we first need to perform the clr transform.

$$\text{clr}(L) = \left(\log(\frac{0.3}{0.1348}), \log(\frac{0.1}{0.1348}), \log(\frac{0.2}{0.1348}), \log(\frac{0.4}{0.1348}), \log(\frac{0.05}{0.1348}), \log(\frac{0.05}{0.1348}) \right)$$

we then multiply this by the transposed contrast matrix and get the solution:

$$\text{ilr}(L) = \left(0.5 \log 6, \frac{\sqrt{2}}{2\sqrt{3}} \log(\frac{3}{4}), \frac{1}{\sqrt{2}} \log 3, \frac{\sqrt{2}}{2\sqrt{3}} \log 8, \frac{1}{\sqrt{2}} \log 8 \right).$$

□

Chapter 4

Further properties of models for compositional data

This chapter describes why a traditional covariance structure is not appropriate for compositions and then suggests several alternatives. The Aitchison normal distribution is then defined and the summary statistics for a compositional sample described. The chapter concludes by suggesting how the Aitchison-normality for a composition can be tested.

In this chapter X is used to denote a D-part random composition with components x_i. Z represents a sample consisting of r compositions denoted Z_r each with components z_{ri}. In the examples I will use $\mathbf{Y} = \begin{pmatrix} 0.3 & 0.2 & 0.4 & 0.1 \\ 0.4 & 0.3 & 0.2 & 0.1 \\ 0.2 & 0.2 & 0.4 & 0.2 \end{pmatrix}$, a sample of three components drawn from some, unknown, random distribution.

4.1 Summarising the covariance structure of a composition

When analysing a D-part composition, Aitchison (1986, pp. 52, 53) warns us against using a 'crude covariance structure and matrix'. The crude structure he describes is the set of all covariances $\kappa_{ij} = (\text{cov}(x_i, x_j) : i, j = 1, \ldots, D)$. There are several reasons why an analysis based on this set could fail. The first problem (a major one) is the negative bias of the unit sum constraint. This bias is that the covariances of a single component of a composition X sum to the negative variance of that component, that is $\Sigma_{j=1}^{D} \kappa_{ij} = -\text{var}(x_i)$. This implies that at least one of the covariances between x_i and another component is negative (excluding the trivial case where x_i is constant). This

implies that the correlations are not free to range over $(-1, 1)$, which can lead to problems when analysing the data set. This problem can easily be noticed when considering the two-part composition, where the correlation between x_1 and x_2 is necessarily -1 regardless of the data. The second problem is the non-existent nature of any simple relationship between the crude covariance matrix of a full composition and that of any sub compositions.

Aitchison (1986, pp. 64–82) introduces several meaningful covariance structures that are appropriate to the space of D-part compositions, S^d. He explains that only the ratios between the components are important as only the relative data are available. Aitchison (1986, p. 65) suggests that we should work with the log-ratios of the components as these are unconstrained. He then goes on to define the covariance structure of a D-part composition as the set of all covariances

$$\sigma_{ij.kl} = \text{cov}(\log(x_i/x_j), \log(x_k/x_l)).$$

The following additional structures (of the covariance of a composition) will now be examined: compositional variation array, the variation matrix, the log-ratio covariance matrix and the centred log-ratio covariance matrix. Aitchison (1986, pp. 76–81) also describes several transformations which can be used to transform the variation matrix, the log-ratio covariance matrix and the centred log-ratio covariance matrix between one another.

4.1.1 Compositional variation array

The first structure defined by Aitchison (1986, pp. 68–71) is termed by him the compositional variation array, but consists of the log-ratio variances and the log-ratio means of the compositional data-set. He defines the compositional log-ratio variance as

$$\tau_{ij} = \text{var}(\log(x_i/x_j)) : i, j = 1, \ldots, D.$$

It is easy to see that $\tau_{ij} = \tau_{ji}$ and $\tau_{ii} = 0$. In passing, the relationship between these compositional log-ratio variances τ_{ij} and the log-ratio covariance $\sigma_{ij.kl}$ can be shown to be

$$\sigma_{ij.kl} = \frac{1}{2}(\tau_{il} + \tau_{jk} - \tau_{ij} - \tau_{kl}).$$

This is shown by Aitchison (1986) in Proposition 4.2 on page 67. The compositional variation array is then defined to consist of: $\tau_{ij}(j > i)$ in the upper triangle of the array, and $\varepsilon_{ij}(i > j)$ (the compositional log-ratio means) in the lower triangle of the array. The compositional log-ratio means are defined

as $\varepsilon_{ij} = \mathrm{E}\{\log(x_i/x_j)\}$. The array is purely a device for displaying both the expected values and the variance of log-ratios. When analysing a sample of compositions, Z, we can use the following, standard, formulae in order to arrive at the values τ_{ij} and ε_{ij}. Finally, $\tau_{ij} = \frac{1}{n}\sum_{r=1}^{D}(\log(z_{ri}/z_{rj}) - \varepsilon_{ij})^2$, and $\varepsilon_{ij} = \frac{1}{N}\sum_{r=1}^{N}\log(z_{ri}/z_{rj})$ where N is the sample size and $n = N - 1$. .

Example 4.1. The compositional variation array for the sample, $\mathbf{Y}$, as at the start of the chapter, is given below.

$$\begin{pmatrix} . & 0.0435 & 0.508 & 0.5353 \\ 0.2310 & . & 0.4023 & 0.3086 \\ -0.0959 & -0.3269 & . & 0.1602 \\ 0.8283 & 0.5973 & 0.9242 & . \end{pmatrix}$$

□

4.1.2 Variation matrix

The above structure is a useful descriptive tool, but better analytical tools can be found. The first more analytical structure is the variation matrix ($\mathbf{T}$). $\mathbf{T}$ is just defined as

$$\mathbf{T} = (\tau_{ij}),$$

that is $\mathbf{T} = (\mathrm{var}(\log(x_i/x_j)) : i, j = 1, \ldots, D)$. The matrix $\mathbf{T}$ is symmetric with zero diagonal. Above the principal diagonal it is identical to the compositional variation array.

Example 4.2. The variation matrix for $\mathbf{Y}$ is given below.

$$\mathbf{T} = \begin{pmatrix} 0 & 0.0435 & 0.508 & 0.5353 \\ 0.0435 & 0 & 0.4023 & 0.3086 \\ 0.508 & 0.4023 & 0 & 0.1602 \\ 0.5353 & 0.3086 & 0.1602 & 0 \end{pmatrix}$$

□

4.1.3 Log-ratio covariance matrix

The next structure described is the log-ratio covariance matrix. This matrix is constructed by calculating the covariances of log-ratios with as common denominator one of the components. The common divisor is usually chosen to be x_D, the last component of the D-part composition. This is merely

convention and conclusions drawn will not differ if a different component is chosen. The (i, j) element of the matrix is the covariance between the log of the ith component over the common denominator and the jth component over the common denominator; that is, the log-ratio covariance matrix is

$$\mathbf{\Sigma} = (\sigma_{ij}) = (\text{cov}(\log(\frac{x_i}{x_D}), \log(\frac{x_j}{x_D})) : i, j = 1, \ldots, d).$$

The relation between the quantities σ_{ij} and $\sigma_{ij.kl}$ is as follows:

$$\sigma_{ij.kl} = \sigma_{ij} + \sigma_{kl} - \sigma_{il} - \sigma_{jk}.$$

This is proved by Aitchison (1986) in Proposition 4.5 on page 77.

Example 4.3. The log-ratio covariance matrix for $\mathbf{Y}$ is

$$\mathbf{\Sigma} = \begin{pmatrix} 0.5353 & 0.3502 & 0.09375 \\ 0.3502 & 0.3086 & 0.03325 \\ 0.09375 & 0.03325 & 0.1602 \end{pmatrix}.$$

□

4.1.4 Centred log-ratio covariance matrix

The final transformation described by Aitchison (1986, pp. 79, 80, 81) is termed the centred log-ratio covariance matrix. This matrix is similar to the log-ratio covariance matrix, the main difference being that common denominator, x_D, is replaced with the geometric mean of the components of X. The matrix is thus

$$\mathbf{\Gamma} = (\gamma_{ij}) = (\text{cov}(\log(\frac{x_i}{g(x)}), \log(\frac{x_j}{g(x)})) : i, j = 1, \ldots, D).$$

The covariance $\sigma_{ij.kl}$ satisfies:

$$\sigma_{ij.kl} = \gamma_{ij} + \gamma_{kl} - \gamma_{il} - \gamma_{jk}.$$

This is shown by Aitchison (1986) in Definition 4.6 on page 79. The matrix $\mathbf{\Gamma}$ is singular.

Example 4.4. The log-ratio covariance matrix for $\mathbf{Y}$ is

$$\mathbf{\Gamma} = \begin{pmatrix} 0.1681 & 0.0548 & -0.1005 & -0.1224 \\ 0.0548 & 0.08498 & -0.0892 & -0.0506 \\ -0.1005 & -0.0892 & 0.139 & 0.0506 \\ -0.1224 & -0.0506 & 0.0506 & 0.1224 \end{pmatrix}.$$

□

The above matrices can be easily transformed into one another using the following formulae:

$$\begin{aligned}
\mathbf{T} \to \mathbf{\Sigma}: &\qquad \sigma_{ij} = \frac{1}{2}(\tau_{iD} + \tau_{jD} - \tau_{ij});\\
\mathbf{\Sigma} \to \mathbf{T}: &\qquad \tau_{ij} = \sigma_{ii} + \sigma_{jj} - 2\sigma_{ij};\\
\mathbf{\Gamma} \to \mathbf{\Sigma}: &\qquad \sigma_{ij} = \gamma_{ij} - \gamma_{iD} - \gamma_{jD} + \gamma_{DD};\\
\mathbf{\Sigma} \to \mathbf{\Gamma}: &\qquad \gamma_{ij} = \sigma_{ij} - \sigma_{i.} - \sigma_{j.} + \sigma_{..};\\
\mathbf{T} \to \mathbf{\Gamma}: &\qquad \gamma_{ij} = \frac{1}{2}(\tau_{i.} + \tau_{j.} - \tau_{ij} - \tau_{..});\\
\mathbf{\Gamma} \to \mathbf{T}: &\qquad \tau_{ij} = \gamma_{ii} + \gamma_{jj} - 2\gamma_{ij}.
\end{aligned}$$

The use of a . in place of an i or j above is used to denote an average. Each of these three matrices enables us to specify a compositional covariance structure. Each of these matrices has advantages and disadvantages. These structures are particularly useful when performing tests of hypotheses about the covariance structure.

4.2 Compositional normal distribution

Aitchison (1986, pp. 112, 113) describes a method of converting the standard multivariate normal distributions on $\mathbb{R}^d$ onto the simplex (S^d). He suggests that we do this by use of one-to-one transformations. There are several preferred transformations. Aitchison (1986, p. 113) suggests that the additive logistic normal class, the multiplicative logistic normal class and the partitioned logistic normal class are the most useful transformed normal distributions. The additive logistic normal class makes use of the additive logistic transformation (see Section 3.1) and is captured by the transformation from $k \in \mathbb{R}^d$ to $X \in S^d$ as follows:

$$x_i = \frac{e^{k_i}}{e^{k_1} + \ldots + e^{k_d} + 1} (i = 1, \ldots, d),$$

$$x_D = 1 - x_1 - \ldots - x_d = \frac{1}{e^{k_1} + \ldots + e^{k_d} + 1}.$$

The inverse function is $k_i = \log(\frac{x_i}{x_D})(i = 1, \ldots, d)$ and the Jacobian is $J = (x_1 \ldots x_D)^{-1}$. Let $N^d(\mu, \Sigma)$ denote the d-dimension normal distribution, with mean vector μ and covariance matrix Σ. A composition, X is said to be distributed as the additive logistic normal distribution, denoted

by $\mathscr{L}^d(\mu, \Sigma)$, when $k = \text{alr}(X)$ is distributed as $N^d(\mu, \Sigma)$. We can use this definition to transform our data X into k and then work with the multivariate normal distribution. In order to work in this manner the reasonableness of a normality assumption on the transformed data would need to be tested. The density function of $\mathscr{L}^d(\mu, \Sigma)$ can be found by transforming the Normal distribution associated with k. The resulting density is:

$$(2\pi)^{-d/2}|\Sigma|^{-1/2}(x_1 \dots x_D)^{-1} \exp(-\frac{1}{2}\{\log(X_{-D}/x_D) - \mu\}'\Sigma^{-1}\{\log(X_{-D}/x_D) - \mu\}).$$

Here X_{-D} is the composition $\mathbf{x}$ with the Dth component removed. It is worth remembering that the support of $\mathscr{L}^d$ is the simplex, S^d.

When the compositional problem involves partitions, then Aitchison (1986, p. 132) suggests the use of either the multiplicative or the partitioned logistic transformations. For X a D-part composition divided into two partitions X_1 and X_2 the multiplicative logistic normal distribution allows us to discuss the relationship between the first and second partitions. A composition X is said to have multiplicative logistic normal distribution $\mathbb{M}(\mu, \Sigma)$ when k, defined by $k_i = \log(\frac{x_i}{1-x_1-\dots-x_i})$ for all $i = 1, \dots, d$ has $N(\mu, \Sigma)$ distribution. The partitioned logistic normal distribution is obtained by performing the following three log-ratio transformations, $y_0 = \log(t/(1-t))$, $y_{1i} = log(s_{1i}/s_{1c})$ and $y_{2i} = \log(s_{2i}/s_{2(D-c)})$ where s_1 is a partition of a composition with c components, s_2 is a partition of the remaining $D - c$ components and t is the sum of the components of the fist partition s_1. If the transformed data are normally distributed then we say that the untransformed data has partitioned log-ratio normality. There seems to be little reference in the literature to either the partitioned logistic normal distribution or the multiplicative logistic normal distribution.

4.3 Summary statistics

Van den Boogaart and Tolosana-Delgado (2013, p. 73) believe that calculating some summary statistics of the composition gives a good overview of the data and of the significance of any results. Van den Boogaart and Tolosana-Delgado begin by defining the compositional mean of a sample data-set (here Z: see page 27), in terms of the centred log-ratio transformation, as $\bar{Z} = \text{clr}^{-1}\left(\frac{1}{N}\sum_{r=1}^{N}\text{clr}(Z_r)\right) = C\left(\exp(\frac{1}{N}\sum_{r=1}^{N}\log(Z_r))\right)$. The closure operation ensures that the compositional mean is actually a composition. They define the 'metric variance' as a global measure of spread, $\text{mvar}(Z) = \frac{1}{N-1}\sum_{n=1}^{N} d_A^2(Z_n, \bar{Z})$. This value is the average squared distance

from the compositional mean. They then define the 'metric standard deviation' $\text{msd}(Z) = \sqrt{\frac{1}{D-1}\text{mvar}(Z)}$.

In order to find μ and Σ for a 'Compositional normal' $\mathscr{L}^d(\mu, \Sigma)$, one does not need the central mean and metric variance; the standard method for finding these measures on the alr-transformed data, y_{ri} is as follows: $\widehat{\mu}_i = N^{-1} \sum_{r=1}^{N} y_{ri}$, $\widehat{\sigma}_{ij} = n^{-1}(\sum_{r=1}^{N} y_{ri}y_{rj} - N\widehat{\mu}_i\widehat{\mu}_j)$, for $i, j = 1, \ldots, d$ where N is the sample size and $n = N - 1$.

4.4 Testing logistic normality of compositional data

Aitchison (1986, pp. 143–148) suggests that we can test if the data are distributed as $\mathscr{L}^d(\mu, \Sigma)$ by testing the log-ratio compositions for multivariate normality.

In order to test the log-ratio normality of compositional data the data are first transformed into $\mathbb{R}^d$ and then standard tests for multivariate normality are applied. Aitchison (1986) suggests that three tests should be used to assess multivariate normality. These tests are termed the marginal test, the bivariate angle test and the radius tests. These test, respectively, the distributions of all the d marginal univariate distributions, the $\frac{d}{2}(d-1)$ bivariate angle distributions and the d-dimensional radius distribution.

After the assumption of logistic normality has been verified, we can freely transform the compositional data into a multi-normal analysis with sample space $\mathbb{R}^d$.

Chapter 5

Hypothesis testing

This chapter begins by explaining an approach to hypothesis testing referred to as a 'lattice of hypotheses'. This approach is not essential when testing hypotheses about a composition, but it is recommended by Aitchison (1986, p. 149). Next the chapter explains how to test a hypothesis about a composition. The tests begin by converting the composition into $\mathbb{R}^d$ and then using standard multivariate tests. The chapter will conclude by testing the hypothesis that the compositions in Tables 1.2 and 2.2 have a common mean and covariance structure.

5.1 Lattice of hypotheses

When analysing a composition, Aitchison (1986, pp. 149, 150) recommends using the 'lattice of hypotheses' approach. This approach starts by testing the most constrained case and then moves on to less constrained cases if this null hypothesis is not accepted. An example of this in comparing two compositions would be to start by testing $\mathrm{H_o} : \mu_1 = \mu_2$ and $\Sigma_1 = \Sigma_2$. If this hypothesis were not accepted, we would then test $\mu_1 = \mu_2$, and also $\Sigma_1 = \Sigma_2$. If we failed to accept either of these hypotheses we would conclude that the samples are unrelated. If both the previous two hypotheses are not rejected our conclusion would be that either $\mu_1 = \mu_2$ or $\Sigma_1 = \Sigma_2$. The lattice for this example is shown in Figure. 5.1

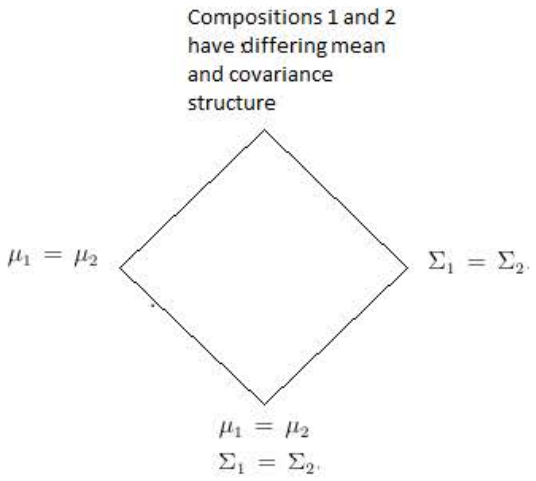


Figure 5.1: Lattice of Hypotheses

5.2 Testing hypotheses about the mean and variation of two compositions

The hypotheses are tested as they would normally be in $\mathbb{R}^d$, using the data after transformation. Aitchison (1986, p. 153) recommends the use of an additive log-ratio transformation, but Pawlowsky-Glahn *et al.* (2015, p. 136) recommend the isometric log-ratio transformation.

A common approach for a problem would be to test using the likelihood test statistic. $L(\Theta|X)$ will denote the likelihood of the parameter Θ for the composition X, with $\widehat{\Theta}_h(X)$ and $\widehat{\Theta}(X)$ being the maximum likelihood estimates for the hypothesis and model respectively. We then let $L_h(X) = L(\widehat{\Theta}_h(X)|X)$ and $L(X) = L(\widehat{\Theta}(X)|X)$ denote the maximised likelihoods under the null hypothesis and the less restricted model. If the maximum likelihood estimates are not available in closed form, then an iterative method can be used to find them. The likelihood ratio test statistic, $R(X) = L(X)/L_h(X)$ can be used, if the distribution under the hypothesis is known, to determine whether the hypothesis is to be accepted. If the exact distribution is unknown then the asymptotic statistic, $Q(X) = 2\log R(X)$ can be compared to the $\chi^2(c)$ distribution, where c represents the difference of the number of parameters under the null hypothesis and the alternative hypothesis.

5.3 Example

Example 5.1. Consider a comparison of the samples of compositions given in Tables 1.2 and 2.2, and refer to these as X and Y respectively. For the purpose of this example I will assume that the compositions are distributed as $\mathscr{L}^2(\mu_1, \Sigma_1)$ and $\mathscr{L}^2(\mu_2, \Sigma_2)$ respectively.

The alr transforms for the data are:

$$\text{alr}(X) = \begin{pmatrix} -0.8535 & -1.6169 \\ -1.3282 & -1.6606 \\ 0.4403 & -0.6312 \\ 0.5160 & -0.4117 \\ -1.3751 & -2.386 \\ -1.2852 & -1.8880 \end{pmatrix}$$

$$\text{alr}(Y) = \begin{pmatrix} -1.0377 & -1.4751 \\ -1.2608 & -1.4741 \\ 0.5771 & -0.3989 \\ 0.5539 & -0.2472 \\ -1.5066 & -1.7375 \\ -1.3837 & -1.4784 \end{pmatrix}.$$

At the first level of testing we test the most restrictive model, that is we test if $\mu_1 = \mu_2$ and $\Sigma_1 = \Sigma_2$. The sample means are: $\widehat{\mu}_1 = (-0.6476131, -1.432444)$ and $\widehat{\mu}_2 = (-0.6763033, -1.135212)$. The sample covariances are: $S_1 = \begin{pmatrix} 0.6631 & 0.5395 \\ 0.5395 & 0.4812 \end{pmatrix}$ and $S_2 = \begin{pmatrix} 0.7911 & 0.5132 \\ 0.5132 & 0.3403 \end{pmatrix}$. Using these matrices we can calculate the pooled covariance matrix estimate: $S = \begin{pmatrix} 0.7271 & 0.5263 \\ 0.5263 & 0.4107 \end{pmatrix}$, and then using the pooled covariance matrix we can calculate the combined sample estimates. $\hat{\mu}_c = (-0.6620, -1.2838)$, the combined sample mean and $S_c = \begin{pmatrix} 0.7273 & 0.5242 \\ 0.5242 & 0.4328 \end{pmatrix}$.

Testing $\mu_1 = \mu_2$ and $\Sigma_1 = \Sigma_2$

The first hypothesis to be tested is the null hypothesis that $\mu_1 = \mu_2$ and $\Sigma_1 = \Sigma_2$. This is tested against the alternative hypothesis that at least one of these equations fails. That is either $\mu_1 \neq \mu_2$ or $\Sigma_1 \neq \Sigma_2$. The test statistic for this hypothesis is 13.69 which is then compared against the upper percentage points of the $\chi^2(5)$ distribution. Since the upper 5 per cent point

of $\chi^2(5)$ is 11.07 we reject the null hypothesis and move on to testing at the next level of the lattice.

Testing $\Sigma_1 = \Sigma_2$

The next hypothesis we test is the hypothesis that $\Sigma_1 = \Sigma_2$. The test statistic for the null hypothesis that $\Sigma_1 = \Sigma_2$ is equal to 6.304152, this is to be compared to the $\chi^2(3)$ distribution. As the upper 5 per cent of the distribution is 7.815 we do not reject the null hypothesis that $\Sigma_1 = \Sigma_2$.

Testing $\mu_1 = \mu_2$

In order to test the null hypothesis $\mathrm{H_o} : \mu_1 = \mu_2$ we need to define two more covariance structures, these are defined as S_{ih} and are found using the following iterative process which is given by Aitchison (1986, p. 155) as a solution to the multivariate Behrens–Fisher problem.

1. $S_{ih} = S_i (i = 1, 2)$.
2. $\mu_h = (N_1 S_{1h}^{-1} + N_2 S_{2h}^{-1})^{-1} (N_1 S_{1h}^{-1} \hat{\mu}_1 + N_2 S_{2h}^{-1} \hat{\mu}_2)$.
3. $S_{ih} = S_i + (\hat{\mu}_i - m_h)(\hat{\mu}_i - m_h)^T$.
4. repeat steps two and three until convergence is reached.

The following covariance structures are found, $S_1 = \begin{pmatrix} 0.7054 & 0.6280 \\ 0.6280 & 0.6663 \end{pmatrix}$ and $S_2 = \begin{pmatrix} 0.8460 & 0.5444 \\ 0.5444 & 0.3580 \end{pmatrix}$. The test statistic is found to be 6.6283 which is compared to the upper 5 per cent of the $\chi^2(2)$ (or exponential) distribution, 5.991. We do not reject the null hypothesis at this level and we conclude that the compositions do have a common mean. That is $\mu_1 = \mu_2$.

In this example the two separate hypotheses that $\mu_1 = \mu_2$, and that $\Sigma_1 = \Sigma_2$ are not rejected, but the hypothesis that $\mu_1 = \mu_2$ *and* $\Sigma_1 = \Sigma_2$ is rejected.

The ilr transformation could have been used in place of the alr transformation and the same conclusions would have been reached. □

Chapter 6

Performing a regression analysis with compositional data

6.1 Introduction

There are two roles for compositions in a regression analysis. Compositions can serve as covariates (that is, predictors) or as response variables. Wherever the composition appears, the method used in order to perform a regression analysis is the same; the composition is transformed into the appropriate $\mathbb{R}^d$ space and a regular analysis is performed using the transformed data. The form of a model with a covariate in S^d is

$$W_i = \alpha + \beta(\text{alr}(X_i))^T + \varepsilon_i$$

or

$$W_i = \alpha + \beta(\text{ilr}(X_i))^T + \varepsilon_i.$$

In this formula W_i represents some real response variable, X_i represents some compositional covariate and $\alpha \in \mathbb{R}$ and $\beta \in \mathbb{R}^d$ represent the model parameters. If an ilr transformation is used then the transformed components can be transformed back into the simplex which results in

$$W_i = \alpha + < X_i, B >_A + \varepsilon_i,$$

where B is the composition created by performing a reverse ilr-transformation on the coefficients β.

Here and in chapter 7 the `compositions` package in R has been used to perform the transformations with the `ilr` and `alr` commands with default

settings. The command `ilr` calls `ilrBase(D)`, which provides a particular matrix V which is then used to produce the ilr transform as described by Van den Boogaart and Tolosana-Delgado (2013, pp. 42–43). The default V is in fact the Helmert matrix displayed in Equation 2.11 of Van den Boogaart and Tolosana-Delgado (2013).

The model with a compositional response is of the form:

$$Y_i = a \oplus (U_i \odot b) \oplus \epsilon_i.$$

In this second model Y_i represents a compositional response, a and b (both compositions) are the parameters and U_i is a real covariate.

For the first step in a regression analysis involving compositions, the components of the composition are transformed to $\mathbb{R}^d$. Once all the variables are in $\mathbb{R}^d$, then a multivariate or univariate regression analysis is performed. After a satisfactory model has been found the model can be converted back into S^d if a suitable transformation can be performed (see Van den Boogaart and Tolosana-Delgado (2013)).This chapter will examine a composition in the role of a covariate and then explore the role of a composition as a response variable.

A discussion of multivariate regression analysis is provided in many regression books, and also by Aitchison (1986, pp. 158–166).

Van den Boogaart and Tolosana-Delgado (2013, pp. 96, 97) suggest that when the response variable is multivariate, one simply deals separately with each of the response variables.

6.2 Compositions as covariates

The model created when a composition is the covariate is one which maps compositions from S^d into a response variable in $\mathbb{R}$.

6.2.1 Plotting the model

A good place to start for any model fitting is to plot how the covariates interact with the response variable(s). Van den Boogaart and Tolosana-Delgado (2013, pp. 104, 105) recommend the use of the 'symbol size and colour' method or the 'pairwise log-ratio plot matrix' technique when plotting compositions. The idea behind the symbol size and colour method is to map the composition on a ternary diagram and then to represent the dependent variable or variables by colour and size at the corresponding compositional points. This gives an idea of how the different components affect the dependent variable. The second technique suggested by Van den Boogaart and

Tolosana-Delgado (2013, pp. 105, 106) is the 'pairwise log-ratio plot matrix'. The array consists of the plots of the response variable against each pairwise log-ratio. This array displays how the response variable reacts to different ratios of the composition.

Example 6.1. For this example and the next, data from Simpson *et al.* (2003) will be used. The compositions consist of the proportions (by weight) consumed by an individual (subject 3) over a 6 day period. The response W is fictitious and has been generated by the following formula:

$W = 2 + < (0.2, 0.26, 0.54), X >_A + \varepsilon$, where $\varepsilon \sim N(0, 0.1)$.

Protein (x_1)	Fat (x_2)	Carbohydrates (x_3)	Random response (W)
0.0177	0.1010	0.8813	4.0353
0.1970	0.1537	0.6494	2.4677
0.5210	0.2395	0.2395	1.6880
0.5000	0.1934	0.3066	1.8371
0.1646	0.1321	0.7033	2.894
0.1565	0.1400	0.7035	2.8535

Plotting the compositions in a ternary diagram and representing the corresponding response variable by way of size and colour yields Figure 6.1.

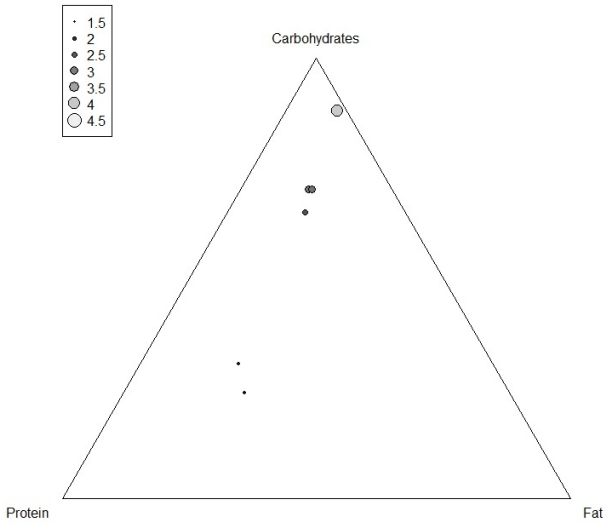


Figure 6.1: Example 6.1; Ternary diagram displaying the covariates for regression analysis

From examining the figure we would hypothesise that our regression analysis would yield a model where a high proportion of carbohydrates yields a higher response.

The response variable can also be plotted against the log-ratios of each component; these plots are displayed in a matrix of plots in Figure 6.2.

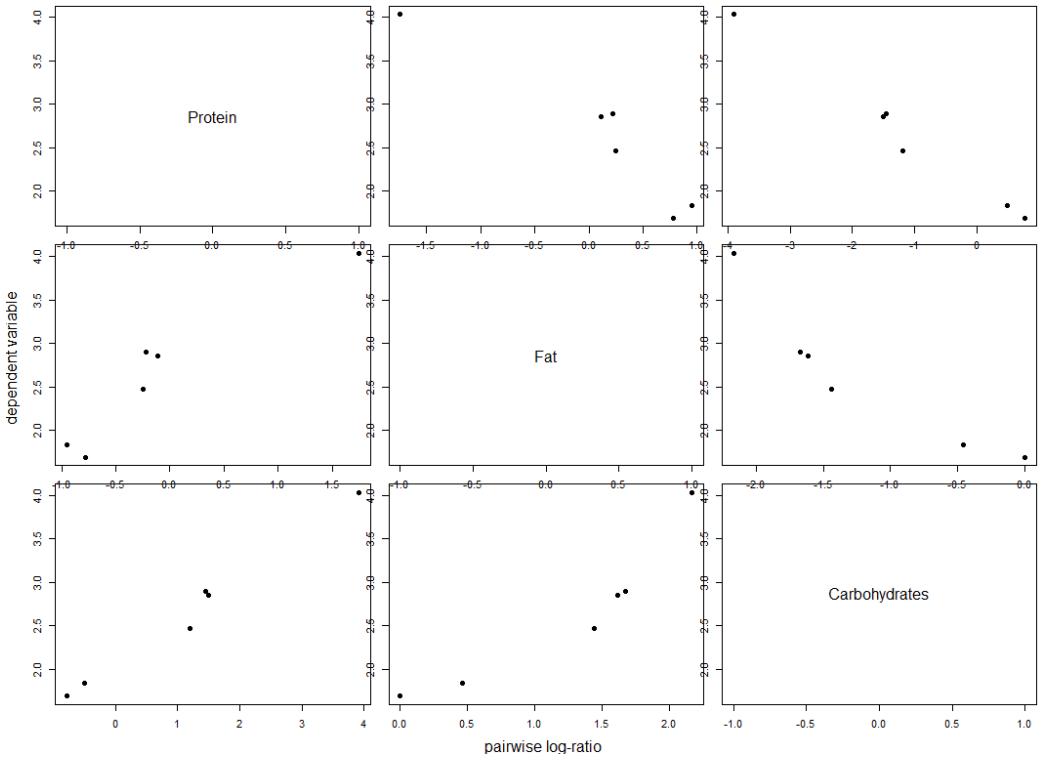


Figure 6.2: Example 6.1; Response variable plotted against all possible log-ratio compositional covariates

□

6.2.2 Fitting the model

The model when the composition is a covariate is

$$W_i = \alpha + \beta \text{alr}(X_i) + \varepsilon_i$$

if the alr transform has been used or

$$W_i = \alpha + \beta \text{ilr}(X_i) + \varepsilon_i$$

if the ilr transform has been used. Van den Boogaart and Tolosana-Delgado (2013, p. 107) show the form of the ilr model when the covariate has been transformed back into the simplex:

$$W_i = a + < b, X_i >_A + \varepsilon_i,$$

$W \in \mathbb{R}$ represents a real response variable, X the compositional covariate. The intercept is a and the composition parameter is b. The model also has an error term, $\varepsilon_i \sim N(0, \Sigma)$. The alr model cannot be transformed back into the simplex, and conclusions are drawn about the log-ratio form. It is important to note that when working with the ilr model transformed into the simplex, that is,

$$W_i = a + < b, X_i >_A + \varepsilon_i,$$

W_i is a single real variable. While when working with

$$W_i = a + \beta \text{alr}(X) + \varepsilon,$$

W_i is in $\mathbb{R}^n$ for some n. For the second model in the previous sentence an ilr transformation may be used in place of an alr transformation.

In order to estimate the parameters the compositions need to be transformed into real space. We can use any transformation to do so, but either an alr or ilr transformation is convenient. After transforming the variables the general model can be written as:

$$W_i = a + \sum_{k=1}^{D-1} \beta_k \text{ilr}_k(X_i) + \varepsilon_i$$

or

$$W_i = a + \sum_{k=1}^{D-1} \beta_k \text{alr}_k(X_i) + \varepsilon_i.$$

The covariate is in unconstrained $\mathbb{R}^d$ space, and thus standard tools can be applied. In the above models, $\text{alr}_k(X_i)$ and $\text{ilr}_k(X_i)$ refer to the k^{th} component of the transformed composition. If the ilr transformation is used and $W \in \mathbb{R}$ then the value for b in the model obtained when the ilr transformed components are transformed into the simplex can be found by use of the inverse ilr transformation; if the alr transformation is used or if $W \in \mathbb{R}^n$ the composition stays in its transformed log-ratio state.

Van den Boogaart and Tolosana-Delgado (2013, p. 108) explain their preference as to why, when metric variances (see page 32) are to be compared, the ilr transformation is preferred to the alr transformation. When using the alr transformation and examining a generalised measure of determination then the geometry of the model is changed and the composition is no longer permutation invariant. Before the ilr transformation was introduced, Aitchison (1986, p. 282) suggested the use of the alr transformation to model the composition; but more recently Aitchison (2008) also stated that the alr transformations are more easily interpretable and that the advantages of this transformation outweigh any supposed drawbacks.

A linear model may be too simple for the problem that is being analysed, in this case it may be necessary to examine a quadratic relationship. Van den Boogaart and Tolosana-Delgado (2013, p. 120) suggest that if we are interested in a quadratic model we need a model with the form:

$$W_i = a + \sum_{j}^{(b_j)} \text{ilr}_j(X_i)_j + \sum_{j} c_{jj} \text{ilr}_j(X_i)_j + \sum_{j<i} c_{jk} 2\text{ilr}_j(X_i)\text{ilr}_k(X_i) + \varepsilon_i.$$

This is equivalent to the following, transformed, model:

$$W_i = a + < b, X_i >_A + < X_i, CX_i >_A + \varepsilon_i.$$

It is important to note that the alr transformation could be used here in place of the ilr transformation in the model in real space. The parameters can still be found with the standard multivariate regression methodologies. The model can be checked using usual test procedures, as in multivariate or univariate regression.

Example 6.2. The data used for this regression analysis can be found in the previous example, along with the formula used to generate the response variables. This example compares an alr and ilr transformed analysis and uses R to perform any linear modelling.

The first model we examine is the one given when an alr transform is performed on the data. The model is of the form $W = a + \beta_1 \text{alr}_1(x) + \beta_2 \text{alr}_2(x)$, with alr_i (as usual) referring to the ith component of an alr transformed composition.

1. alr transformed composition

Model in $\mathbb{R}$	$: 2.0728 - 0.5097 \log(x_1/x_3) + 0.0167 \log(x_2/x_3)$
Bias adjusted sum of residuals squared	$: 0.977$

We can also perform a transformation with one of the values of the composition removed. The model below is a linear model of the data with the second component of the composition removed.

2. alr transformed composition (with 2nd value removed)

Model in $\mathbb{R}$	$: 2.0610 - 0.50210 \log(x_1/x_3)$
Bias adjusted sum of residuals squared	$: 0.9827$

The final model we will examine is a model on the ilr transformed composition. This model will be of the form $W_i = a_i + \beta_1 \text{ilr}_1(x) + \beta_2 \text{ilr}_2(x)$.

3. ilr transformed composition

Model in $\mathbb{R}$	$: 2.0728 - 0.6039\text{ilr}_1(x) - 0.3721\text{ilr}_2(x)$
Compositional form of the model	$: 2.0728 + < (0.1846, 0.3124, 0.5031), X >_A$
Bias adjusted sum of residuals squared	$: 0.977$

Although the two methods produce models with identical results, the ilr transformation can be back-transformed into the simplex. □

For multiple hypothesis testing there may be a hierarchy of hypotheses; Aitchison (1986, p. 162) once again recommends a lattice approach, starting with the least constrained models and moving on to the more constrained ones. He also suggests the use of a standard F-test with the test statistic: $\frac{(R_h - R_m)/(p_m - p_h)}{R_m/(N - p_m)}$, where N represents the number of different mixtures, R_h and R_m are the residuals of the hypothesis and model respectively, and p_m and p_h are the number of parameters under the model and hypothesis. The test statistic should be tested against the upper percentage points of the $F(p_m - p_h, N - p_m)$ distribution.

6.3 Compositions as response variables

Here a composition plays the role of response variable, and the model has real covariates and produces from them a composition in S^d.

6.3.1 Plotting the model

Once again it may be prudent to explore a potential model via some form of visualisation. Van den Boogaart and Tolosana-Delgado (2013, p. 125-129) suggest methods of exploration. They begin with the case where the covariates are continuous, where one can generate an array of plots. For each plot the covariate is plotted along the x-axis against a transformed composition (the composition could be transformed via clr, alr, ilr or pairwise log-ratio transformations). Where the covariates are discrete we can denote the possibilities with different colours or symbols and plot the dependent

composition on a ternary diagram or some other visual display with each associated variable. I now provide an example of such a visual representation.

Example 6.3. For this example and the next, the following fictitious data will be used. The following formula was used to generate this data, $Y_i = (0.3, 0.2, 0.5) \oplus U_i \odot (0.4, 0.3, 0.3) \oplus \epsilon_i$ where $\epsilon \sim \mathscr{L}(0, I_2)$. The simulated data are as follows:

Sample(i)	U_i	Y_i
1	4	(0.5726, 0.1216, 0.3057)
2	5	(0.6371, 0.1061, 0.2568)
3	9	(0.8459, 0.0439, 0.1102)
4	8	(0.8082, 0.0556, 0.1362)
5	3	(0.5081, 0.1399, 0.3519)
6	12	(0.9285, 0.0203, 0.0512)
7	11	(0.9079, 0.0265, 0.0656)
8	8	(0.8129, 0.0524, 0.1346)

The resulting tertiary diagram, is Figure 6.3, which displays the response variables with symbol size representing the size of the covariate.

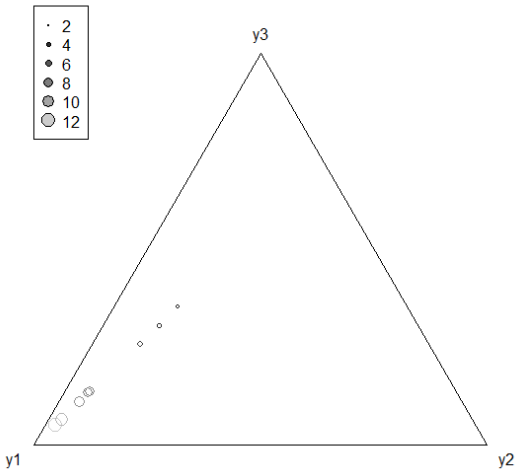


Figure 6.3: Ternary diagram displaying the response for Example 6.3

From Figure 6.3 we see that when the value of U increases we expect the proportion of the component, y_1 to increase.

□

6.3.2 Fitting the model

Van den Boogaart and Tolosana-Delgado (2013) define the model on the simplex:

$$Y_i = a \oplus (U_i \odot b) \oplus \epsilon_i.$$

In this model, a and b are constant compositions, and $Y_i \in S^d$ is the response variable. U_i is a real covariate and ϵ_i is the error term, with $\epsilon_i \sim \mathscr{L}(0, \Sigma)$. The model must be transformed into $\mathbb{R}^d$ in order to treat it as a normal multivariate response model. Van den Boogaart and Tolosana-Delgado (2013, p. 129) recommend an ilr transformation, which will be described here, but Aitchison (1986, pp. 158,159) recommends an alr transformation.

The ilr transformation yields

$$\text{ilr}(Y_i) = \text{ilr}(a) + U_i\text{ilr}(b) + \varepsilon_i,$$

with $\varepsilon_i \sim N(0, \Sigma)$. The parameters a and b can now be estimated via standard multivariate regression. The alr transformation could have been used in place of the ilr transformation. For notational ease I will define α as equal to $\text{ilr}(a)$ and β as $\text{ilr}(b)$. The estimators for α and β are the same as the usual least squares estimates. The model can be extended to handle cases where there are multiple covariates, discrete or continuous. This expansion is fairly simple and extra terms are added to the model where appropriate. The model,

$$Y_i = a \oplus U_i^1 \odot b \oplus \ldots \oplus U_i^N \odot b_N \oplus \varepsilon_i$$

with N representing the number of covariates, can be easily transformed to

$$\text{ilr}(Y_i) = \text{ilr}(a) + U_i\text{ilr}(b_1) + \ldots + U_i^N\text{ilr}(b_N) + \epsilon_i.$$

The terms in the expression are defined as they have been previously. The alr transformation could also have been used above in place of the ilr here.

Model adequacy can be tested via the standard methodologies for multivariate regression analysis. The significance of each variable can be tested with the use of an analysis of variance table. We can use the table, for instance, to test the null hypothesis that $\text{ilr}(b_i) = 0$ against the alternative hypothesis that $\text{ilr}(b_i) \neq 0$. The p-values for the test can be found in the anova table if the software R is used. Any non significant covariates should be removed. Van den Boogaart and Tolosana-Delgado (2013, pp. 130-133) suggest that we should test each variable as the last one in the model if R is being used.

Example 6.4. The alr and ilr methods are used below to fit models to the simulated data of Example 6.3. The practice followed is to transform the

response variable into $\mathbb{R}^2$ and then model the two responses separately. The model parameters found through this process can then be transformed back into the simplex with the inverse of the transformation used on Y. The models are as follows:

The first model examined is the one produced via an alr transform. The form of the model in $\mathbb{R}^d$ is $\text{alr}(Y) = \alpha + U\beta$, where α and β are of the same dimensions as $\text{alr}(Y)$ when $U \in \mathbb{R}$, as it is here.

1. alr transformed

$$\text{Model in } \mathbb{R}^2 : \text{alr}(Y) = \begin{pmatrix} -0.4940 \\ -0.9106 \end{pmatrix} + x \begin{pmatrix} 0.2833 \\ -0.0005 \end{pmatrix}$$

$$\begin{matrix} \text{Compositional} \\ \text{form of the model} \end{matrix} : Y = (0.303, 0.200, 0.497) \oplus U \odot (0.399, 0.300, 0.301)$$

The second model examined is the model produced via an ilr transformation on the response variabe. The model found in $\mathbb{R}^d$ is $\text{ilr}(Y) = \alpha + x\beta$, where α and β are of the same dimensions as $\text{ilr}(Y)$ when $x \in \mathbb{R}$.

2. ilr transformed

$$\text{Model in } \mathbb{R}^2 : \text{ilr}(Y) = \begin{pmatrix} -0.2946 \\ 0.5734 \end{pmatrix} + U \begin{pmatrix} -0.2007 \\ -0.1154 \end{pmatrix}$$

$$\begin{matrix} \text{Compositional} \\ \text{form of the model} \end{matrix} : Y = (0.303, 0.200, 0.497) \oplus x \odot (0.399, 0.300, 0.301)$$

The ilr and alr models both capture the data almost perfectly and as expected produce the same results when transformed back into the simplex. □

6.4 Remarks

6.4.1 The denominator of the alr transform

A natural question that may arise is how the choice of any one particular component as denominator in the alr transformation affects the model. Although Aitchison (1986, p. 142) assures us that the choice of this component will not have any effect there may still be some doubt in a practitioner's mind. When the composition takes the role of a covariate, as in Section 6.2, it is

simple to show that the choice of component is irrelevant. The reader can verify that

$$\alpha + \beta_1 \log(\frac{x_1}{x_3}) + \beta_2 \log(\frac{x_2}{x_3}) = \alpha + \beta_1 \log(\frac{x_1}{x_2}) - (\beta_1 + \beta_2) \log(\frac{x_3}{x_2}).$$

When the composition is the response variable, as in 6.3, it is also easy to show that the choice of denominator is irrelevant. The reader can once again verify that

$$\begin{pmatrix} \log(y_1/y_3) = \alpha_1 + \beta_1 x \\ \log(y_2/y_3) = \alpha_2 + \beta_2 x \end{pmatrix} = \begin{pmatrix} \log(y_1/y_2) = (\alpha_1 - \alpha_2) + (\beta_1 - \beta_2)x \\ \log(y_3/y_2) = -\alpha_2 - \beta_2 x \end{pmatrix}.$$

6.4.2 Compositions as both the response variable and the covariate

The next issue to discuss might be how the model is formed if compositions fill both the roles of response and covariate. We transform any composition from S^d into $\mathbb{R}^d$ and then perform the modelling process as before. The model, in S^d will be of the form:

$$Y = \alpha_c \oplus (\beta_c \odot X) \oplus \varepsilon$$

which is a combination of the previous two cases. The model in $\mathbb{R}^d$ is

$$\text{alr}(Y) = \alpha_{ar} + \beta_{ar}\text{alr}(X)$$

using an alr transformation and

$$\text{ilr}(Y) = \alpha_{ir} + \beta_{ir}\text{ilr}(X)$$

when using an ilr transformation. Once the data have been transformed to unconstrained space then we use normal multivariate linear modelling techniques to create a model of the data.

6.4.3 A comparison of the model structures

When a composition plays any role in the regression analysis, a model can be written in $\mathbb{R}$ or in the simplex, except when the composition plays the role of a covariate and the alr transformation has been used. In this case the transformed components cannot be transformed back into the simplex after a model has been found. The following models can be obtained when the composition plays the role of a covariate and are seen in Section 6.2:

$$W_i = a + < b, X_i >_A + \varepsilon_i$$

$$W_i = \alpha + \beta \text{ilr}(X_i) + \varepsilon_i.$$

Section 6.3 describes the models found when the response is a composition. A model in the simplex and one in real space are as follows:

$$Y_i = a \oplus (U_i \odot b) \oplus \epsilon_i$$

$$\text{ilr}(Y_i) = \text{ilr}(a) + U_i \text{ilr}(b) + \varepsilon_i.$$

Finally, 6.4.2 describes two models for when compositions are found both in the response and the covariates. The models, seen below, are defined in the simplex and in real space respectively:

$$Y = \alpha_c \oplus (\beta_c \odot X) \oplus \epsilon$$

$$\text{ilr}(Y) = \alpha_{ir} + \beta_{ir} \text{ilr}(X) + \varepsilon_i.$$

The alr transformation could have been used in place of the ilr transformation in all of the above.

Chapter 7

Examples

In this chapter I discuss four examples that show how a regression analysis can be performed with compositional data. The examples below serve as demonstrations of compositional techniques, even if some of the results are negative. Unless otherwise indicated, the F-statistic referred to here is the F-statistic of a test of the regression model with all explanatory variables included against one with only an intercept term and random variation. I will also refer to the model with the response variable explained by only an intercept and random variation as the null model.

The first three examples in this chapter include compositions as covariates with a real variable as a response. The fourth example examines the response as a composition with real variables as covariates.

7.1 Example 1: Neoplasms explained by proportions in different age groups

This example will consider what, if any, effect the age structure of a population has on the number (per 100 000) of discharges of patients with a neoplasm. (Here discharge means that a patient leaves due to finalisation of treatment, signs out against hospital advice, transfers to another health care institute or dies.) The data used will be the 2004 population statistics for 24 of the EU members and the rates per 100 000 of discharges of in-patients with a neoplasm for those countries in the same year. All of the data were taken from "`ec.europa.eu/eurostat`". I have provided a summary of the data in Table 7.1, and this summary is the only data I use. A similar analysis was performed by Hron *et al.* (2012), who used the number of neoplasm discharges in 2007 and the population structure for 2008.

The proportions of the population in different age ranges provide the

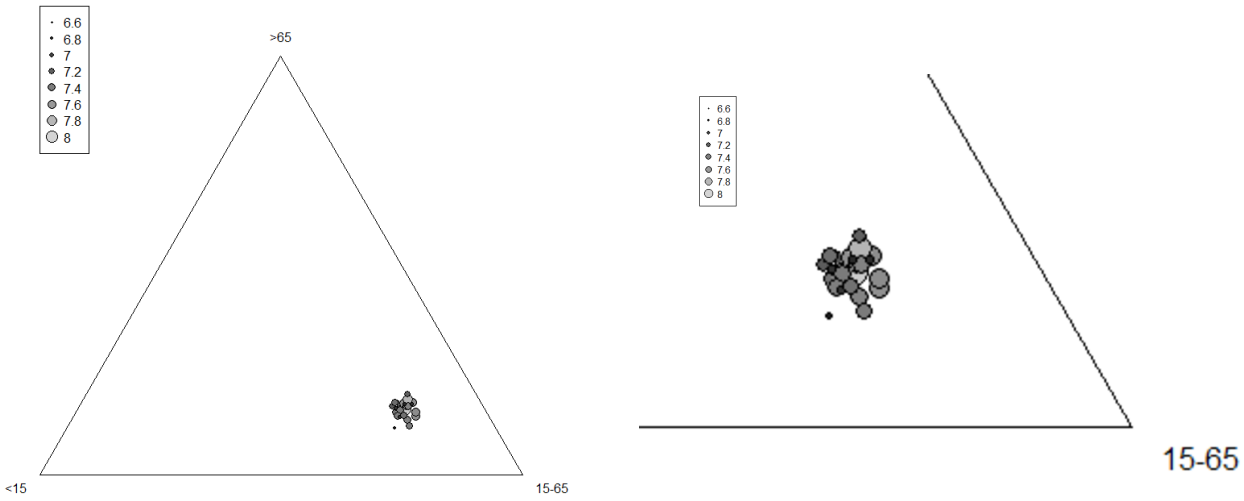
Table 7.1: 24 EU countries, their population age percentages and the number of neoplasm discharges (per 100 000) for 2004.

Country	Age: <15	Age: 15–65	Age: >65	neoplasms (per 100 000)
Austria	16.3	68.1	15.5	2823.8
Belgium	17.3	65.6	17.1	1223.1
Bulgaria	14.1	68.6	17.3	1827.6
Croatia	16.1	66.7	17.1	1852.5
Czech	15.2	70.9	14.0	1856.9
Denmark	18.9	66.2	14.9	1520.1
Estonia	15.8	67.9	16.2	1630.2
Finland	17.6	66.8	15.6	1856.4
France	18.8	65.0	16.2	1299.3
Germany	14.7	67.3	18.0	2401.1
Ireland	20.9	68.0	11.1	818.4
Italy	14.2	66.6	19.2	1330.4
Latvia	15.5	68.3	16.2	1686.3
Lithuania	17.6	67.0	15.4	1611.7
Luxembourg	18.7	67.3	14.0	1642.9
Netherlands	18.5	67.6	13.8	980.8
Poland	17.2	69.8	13.0	1704.4
Portugal	16.0	67.1	16.9	920.3
Romania	17.5	68.3	14.1	1505.7
Slovakia	17.6	70.8	11.6	1608.2
Slovenia	14.6	70.4	15.0	1814.4
Spain	14.5	68.6	16.8	932.6
Sweden	17.8	65.0	17.2	1462.8
UK	18.3	65.8	15.9	977.6

covariate for the model, and the number of discharges the response variable. I will use X to denote the composition consisting of the proportions of ages with x_1 referring to the < 15 component, x_2 the 15–65 component and x_3 the > 65 component. W is used to represent the number of inpatient discharges per 100 000 of the population.

During the analysis I found that the logarithm of W is better captured by a linear model in R than was W itself. My analysis uses as the response variable the logarithm of the neoplasm discharge rates per 100 000 of the population, that is $\log(W)$.

I will start by analysing the data via a ternary diagram using the 'symbol size and colour' method as well as a matrix of scatter-plots representing the response variable against the log-ratios. The images include a point for each country included in the analysis. In the ternary diagram the size of each point corresponds to the relative size of the log of the number of discharges in patients with neoplasms. The scatter-plots plot the rate per 100 000 of discharges against all the log-ratios $\log(x_i/x_j)$, which results in some redundancy in Figure 7.2. The numerator of the log-ratio is the component common to the row and the denominator the component common to the column. These diagrams can be seen in Figure 7.1 and Figure 7.2 respectively.



(a) Full ternary diagram

(b) Zoomed in on cluster

Figure 7.1: Ternary diagram showing the effects of the age structure of a country on the log of the number of neoplasm discharges.

Figure 7.1 shows that our data consist mainly of individuals aged 15 to 65, which is expected. The only point of interest seems to be that of Ireland, the

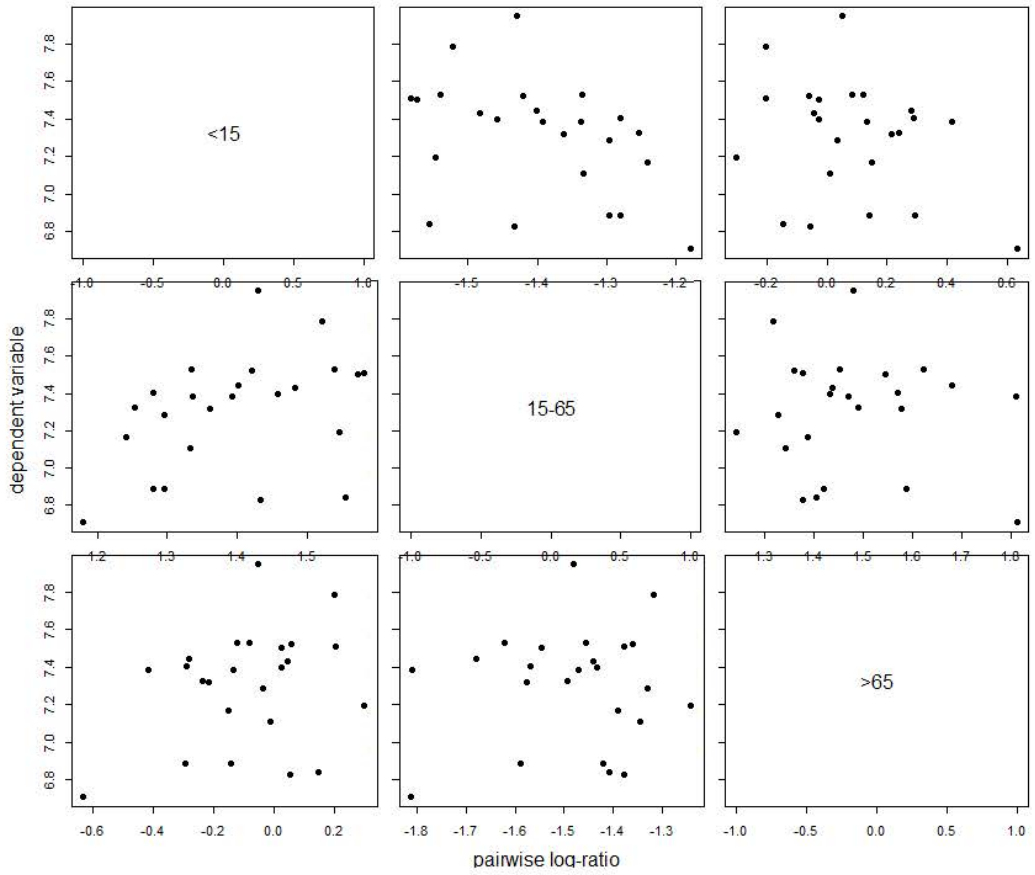


Figure 7.2: An array representing the log of the rate per 100 000 of neoplasm discharges against all possible log-ratios of components

smallest point. This point has the highest proportion of young individuals. Figure 7.2 seems to show that the log-ratios of the age structure of a country have no systematic effect on the number of neoplasms.

A linear model for the effects of the composition on the neoplasm rates will now be built using R and displayed below. I start by modelling the log of the neoplasm rates with an alr-transformed composition (with x_3 as the denominator) of age proportion, and then use an ilr-transformation. The row vector (a^1, a^2) denotes the alr-transformed composition and (s^1, s^2) an ilr-transformed composition, as provided by the `ilr` transformation from the `compositions` package with default settings.

7.1.1 Full model

The alr-transformed model: $\log(W_i) = \alpha + \beta_1 a_i^1 + \beta_2 a_i^2 + \varepsilon_i$.

coefficients:

	Estimate	t value	$\Pr(>\|t\|)$
α	5.7828	5.022	5.69×10^{-5}
β_1	−1.0401	−1.887	0.073
β_2	1.0866	1.353	0.190

The F statistic, with two and 21 degrees of freedom, is 1.899 and the p-value of the F-test is 0.1745. This statistic tells us that we would not require a model of this form (with covariates), and that a model where the neoplasms are not affected by the population structure is adequate. The results also indicate that the a^2 covariate is not very significant, and this suggests removing it. The intercept is seen to be highly significant and the effects of a^1 marginally significant.

The ilr transformed model: $\log(W_i) = \alpha + \beta_1 s_i^1 + \beta_2 s_i^2 + \varepsilon_i$.

coefficients:

	Estimate	t value	$\Pr(>\|t\|)$
α	5.78281	5.022	5.69×10^{-5}
β_1	1.50379	1.630	0.118
β_2	−0.05686	−0.105	0.917

The F statistic is the same as that of the above alr model, with the same number of degrees of freedom. We would draw the same conclusions as above regarding the suitability of the population composition to model the number of neoplasms. The intercept is the only parameter which is significant, and it is highly significant.

After modelling the data with the full composition I conclude that the model with both covariates is not needed in either case. The next step in the process would be to remove one of the components from the composition and test whether the new composition is more appropriate. A prudent approach would be to test all three models with one of the components removed.

7.1.2 Model with the first component removed

In this and the next two subsections I consider a two-part composition where one component from the composition is removed before closure is applied.

The purpose is to consider whether a single log-ratio is sufficient for the model. I start with the alr-transformation. In this subsection a^1 will refer to the alr transformed data (with the second of the remaining components chosen as the denominator of the ratio, originally x_3) of the composition formed when the first component has been removed. Similarly s^1 will refer to the ilr-transformed data for the age structure with the first component removed.

The alr-transformed model with x_1 removed: $\log(W_i) = a + b_1 a_i^1 + \varepsilon_i$.

coefficients:

	Estimate	t value	$\Pr(>\|t\|)$
a	7.6075	11.511	8.8×10^{-11}
b_1	−0.2046	−0.461	0.65

The ilr transformed model with x_1 removed: $\log(W_i) = a + b_1 s_i^1 + \varepsilon_i$.

coefficients:

	Estimate	t value	$\Pr(>\|t\|)$
a	7.6075	11.511	8.8×10^{-11}
b_1	0.2894	0.461	0.65

The two transformations produce the same F statistic with the same degrees of freedom, thus they will be analysed together. The F statistics are both 0.2122 with 1 and 22 degrees of freedom. This produces a p-value of 0.6496, which indicates that we would not choose this model over one with no covariates. This tells us that we would likely not accept that the number of neoplasms should be modelled via a composition consisting of the 15–65 and > 65 age groups. The model does not appear superior to the null model.

7.1.3 Model with the second component removed

For this section the 15–65 age group has been removed from the composition. The symbol a^2 will represent the alr-transformed composition and s^2 the ilr-transformed composition. The alr-transformed model is displayed first, followed by the ilr-transformed model. The alr model with x_2 removed:

$\log(W_i) = c + da_i^2 + \varepsilon_i$.

coefficients:

	Estimate	t value	Pr($>|t|$)
c	7.33858	110.930	$< 2 \times 10^{-16}$
d	−0.40468	−1.377	0.182

The ilr model with x_2 removed: $\log(W_i) = c + ds_i^2 + \varepsilon_i$.

coefficients:

	Estimate	t value	Pr($>|t|$)
c	7.33858	110.930	$< 2 \times 10^{-16}$
d	0.57230	1.377	0.182

The alr and ilr models once again produce the same F statistic in $\mathbb{R}$. The F statistic with 1 and 22 degrees of freedom is 1.896. This produces a p-value of 0.1824, indicating that this model is not superior to one with no covariates. We would once again conclude that the neoplasms are not affected by this composition of age structures. The model indicates no association between the number of neoplasm discharges and the composition representing the proportion of the younger(< 15) to the older(> 65) members of a population.

7.1.4 Model with the third component removed

In this section the > 65 age group has been removed from the composition, leaving the composition to represent only the proportion of younger individuals to middle aged ones. Once more a will be used to represent the alr-transformed composition and s the ilr-transformed composition.

The alr-model with x_3 removed: $\log(W_i) = q + ua_i + \varepsilon_i$.

coefficients:

	Estimate	t value	Pr($>|t|$)
q	5.8760	8.164	4.21×10^{-08}
u	−1.0226	−1.991	0.059

The ilr model with x_3 removed: $\log(W_i) = q + us_i + \varepsilon_i$.

coefficients:

	Estimate	t value	$\Pr(>\|t\|)$
q	5.8760	8.164	4.21×10^{-08}
u	1.4462	1.991	0.059

These models each have an F-statistic of 3.966 on one and 22 degrees of freedom. This statistic yields a p-value of 0.059, which indicates that we could accept that the number of neoplasms are related to the composition consisting of individuals aged < 15 and 15–65. The intercept term is, as usual, highly significant and the covariate is seen to have some effect on the response variable. Transforming the ilr transformed component and the explanatory variables of the model into the simplex yields the following:

$$\log(W) = 5.8760 + < (0.1145, 0.8855), X_{-3} >_A + \varepsilon,$$

where X_{-3} represents the composition with the 3rd component removed in S^2 with components x_1^{-3} and x_2^{-3}. The alr model yields the following:

$$\log(W) = 5.8760 - 1.0226 \log(\frac{x_1^{-3}}{x_2^{-3}}) + \varepsilon.$$

7.1.5 Conclusions

The model indicates that only the ratio of young to middle-aged individuals has some importance to the rate of discharges of inpatients due to a neoplasm It indicates that the proportion of older patients plays little or no role in the neoplasm rate.

7.2 Example 2: Firework mixtures

In this section I will attempt to model the effects which a mixture of five constituents has on the vorticity and brilliance of a firework. The data I use come from an experiment used by Aitchison (1986, pp. 291–293), and I seek to answer the following two questions: do the last two components have any effect on brilliance, and do the first two components have any effect on vorticity? The data used are available through the `data(Firework)` command in R if the `compositions` package has been loaded. Aitchison (1986) explains that the first two components of the mixture, x_1 and x_2,

act as the illuminating components of the firework. The third component, x_3, acts as an accelerator and the final two components, x_4 and x_5, act as binders for this accelerator. The brilliance will be referred to as W and the vorticity as U. I will first examine the brilliance of the fireworks and then the vorticity.

7.2.1 Brilliance

I start by analysing a matrix of scatter-plots of the data in order to suggest relations of interest. These scatter-plots can be seen in Figure 7.3.

The scatter-plots seem to indicate a positive influence of the log ratio of the second and final components (x_2/x_5), on the brilliance of the firework. The alr transformation of the full model, with x_5 chosen as the denominator and a^1 to a^4 representing the elements of the alr-transformed composition, will be referred to as Model A. The model takes the form: $W_i = \alpha + \beta_1 a_i^1 + \beta_2 a_i^2 + \beta_3 a_i^3 + \beta_4 a_i^4 + \varepsilon_i$.

coefficients:

	Estimate	t value	$\Pr(>\|t\|)$
α	14.6802	28.101	$< 2 \times 10^{-16}$
β_1	0.1796	0.281	0.77967
β_2	2.1315	3.331	0.00134
β_3	−0.9111	−1.424	0.15854
β_4	0.9315	1.456	0.14956

Model A seems to suggest that the brilliance is not much affected by a^1, a^3 and a^4. This may suggest that the components x_1, x_3 and x_4 or x_5 do not affect the brilliance of the firework. This tells us that the first component added for brilliance may not affect the brilliance after all. This model does, however, capture the data very well with an F-statistic of 3.831 on four and 76 degrees of freedom. This statistic produces a p-value of 0.0069 which suggests that we can be fairly certain that at least one of the covariates are necessary to model the brilliance.

The ilr-transformed model, with i^1 to i^4 representing the ilr-transformed components will be referred to as model B. This model has the form: $W_i = \alpha + \beta_1 i_i^1 + \beta_2 i_i^2 + \beta_3 i_i^3 + \beta_4 i_i^4 + \varepsilon_i$.

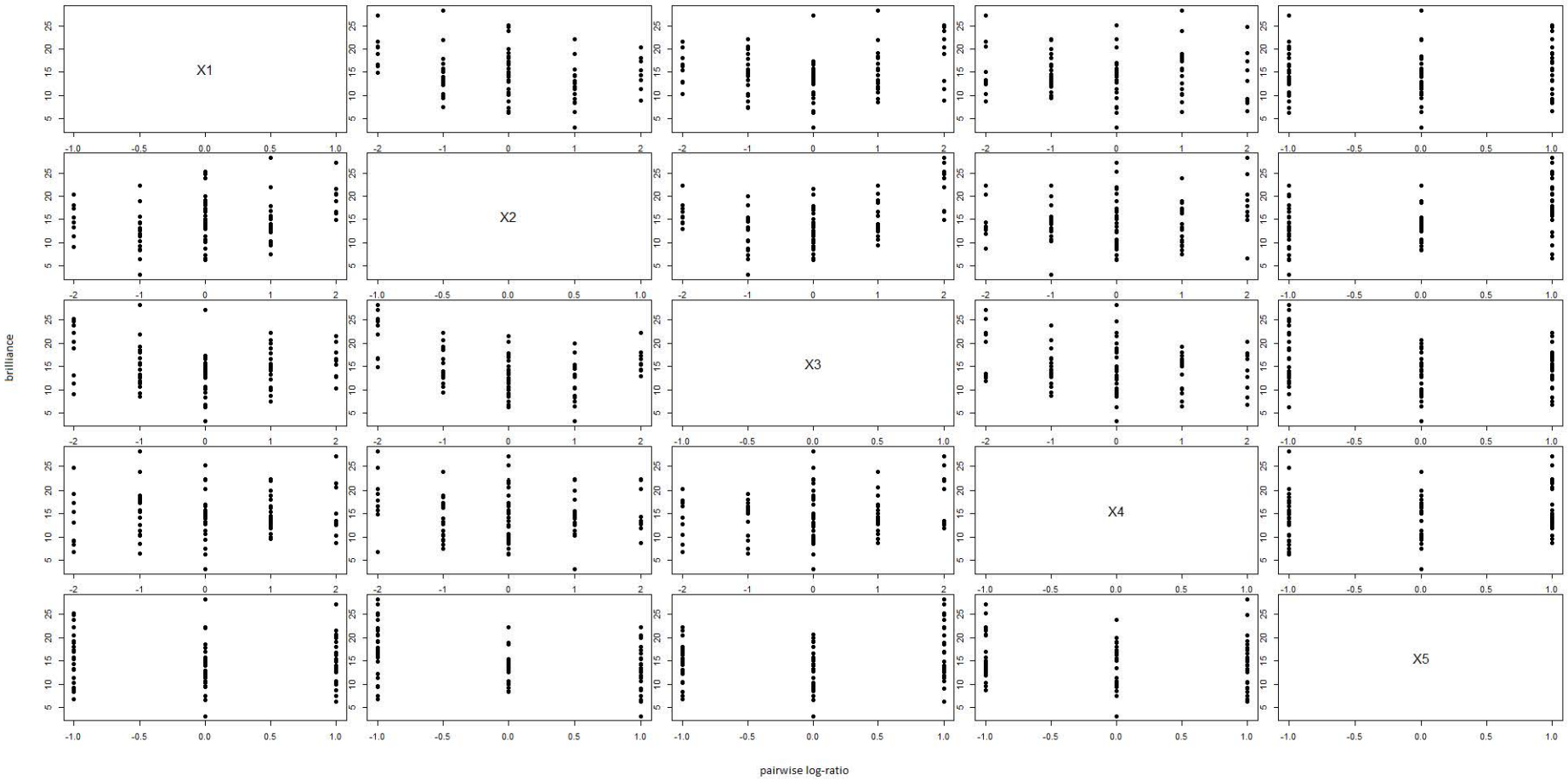


Figure 7.3: An array representing the brilliance of the firework against all possible log-ratios of the mixture composition

coefficients:

	Estimate	t value	$\Pr(>\|t\|)$
α	14.6802	28.101	$< 2 \times 10^{-16}$
β_1	1.3802	2.157	0.0342
β_2	−1.6874	−2.637	0.0101
β_3	0.4025	0.629	0.5311
β_4	−2.6067	−1.822	0.0724

Model B seems as appropriate as model A above. No conclusions can be drawn from model B about the appropriateness of any components as the ilr transformed components are not as easily interpretable as those of the alr.

Model with components 4 and 5 removed

Aitchison (1986) tested whether the last two components of the composition were appropriate in modelling the brilliance of the firework. Both the ilr and alr transformations can be used to test this. I will now model the brilliance of the firework against a composition with components four and five removed. The alr transformed model, with (a^1, a^2) representing the alr transformed components: $W_i = \alpha + \beta_1 a_i^1 + \beta_2 a_i^2 + \varepsilon_i$ is as below.

coefficients:

	Estimate	t value	$\Pr(>\|t\|)$
α	14.6802	27.797	$< 2 \times 10^{-16}$
β_1	−0.2870	−0.543	0.5883
β_2	1.6648	3.152	0.0023

The ilr transformation, with (s^1, s^2) representing the ilr transformed components, $W_i = \alpha + \beta_1 i_i^1 + \beta_2 i_i^2 + \varepsilon_i$ is:

coefficients:

	Estimate	t value	$\Pr(>\|t\|)$
α	14.6802	27.797	$< 2 \times 10^{-16}$
β_1	1.3802	2.134	0.0360
β_2	−1.6874	−2.609	0.0109

Both models are more appropriate than the full model, when observing the brilliance of a firework, as judged by the p-values. The F-statistic is the same for both models and is equal to 5.679 on two and 78 degrees of freedom, the p-value associated with testing if a model is superior to the null model is 0.004981. This p-value is lower than that of the previous model. This suggests that the model for the brilliance of the firework without the binding components (x_4 and x_5) included is more appropriate than the one including them. The model with only components 4 and 5 removed is not ideal though, as the alr model suggests that the β_1 is not significantly different from 0. The ilr model makes no such suggestion and if we were only using it we would be suitably pleased with this model. The ilr model does, however, allow us to transform the composition back into the simplex. The ilr model with a compositional input $X_{-(4,5)}$ is:

$$Y = 14.68 + < (0.119, 0.840, 0.0401), R >_A + \varepsilon,$$

the alr model is as follows:

$$Y = 14.68 - 0.287 \log(\frac{x_1^{-(4,5)}}{x_3^{-(4,5)}}) + 1.665 \log(\frac{x_2^{-(4,5)}}{x_3^{-(4,5)}}) + \varepsilon.$$

Model with first and fifth components

The alr transformation suggests that a more appropriate model than one consisting of x_1, x_2 and x_3 might exist. Model A seems to suggest that β_1, β_3 and β_4 are not significant.

The alr transformation for the model with only x_2 and x_5 is $W_i = \alpha + \beta_1 \log(\frac{x_2}{x_5}) + \varepsilon_i$ and is analysed below.

coefficients:

	Estimate	t value	$\Pr(>\|t\|)$
α	14.6802	27.885	$< 2 \times 10^{-16}$
β_1	2.1315	3.306	0.00143

This model seems superior to the null model, with an F-statistic of 10.93 on one and 79 degrees of freedom. The p-value associated with testing the model against the null model is 0.001426, indicating that x_2 and x_5 are important when modelling the brilliance of a firework. The alr transformation is more easily interpretable and can be used to improve a model. However, when testing whether one specific model is superior to any other specific model the ilr model could be used just as efficiently.

7.2.2 Vorticity

Aitchison (1986) wanted to test if the vorticity of a firework was dependent on the first two components of the mixture. I will thus use alr and ilr transformations to test whether x_1 and x_2 can reasonably be excluded from the model. The full model using the alr transformation is $U_i = \alpha + \beta_1 a_i^1 + \beta_2 a_i^2 + \beta_3 a_i^3 + \beta_4 a_i^4 + \varepsilon_i$, with coefficients as follows:

	Estimate	t value	$\Pr(>\|t\|)$
α	14.5099	25.873	$< 2 \times 10^{-16}$
β_1	1.8130	2.640	0.0101
β_2	1.5426	2.246	0.0276
β_3	1.4611	2.127	0.0366
β_4	1.2537	1.825	0.0719

The full model using an ilr transformation is $U_i = \alpha + \beta_1 i_i^1 + \beta_2 i_i^2 + \beta_3 i_i^3 + \beta_4 i_i^4 + \varepsilon$, with coefficients as follows:

	Estimate	t value	$\Pr(>\|t\|)$
α	14.5099	25.873	$< 2 \times 10^{-16}$
β_1	-0.1912	-0.278	0.782
β_2	-0.1769	-0.258	0.797
β_3	-0.3047	-0.444	0.659
β_4	-6.7869	-4.419	3.25×10^{-05}

These two models appear to be superior to the null model. The F-statistic for both of the models is 4.967 on four and 76 degrees of freedom. The p-value for a test of these models against the null model is 0.001305, which indicates that we are confident that these models are superior to the null model . The alr model implies that all of the components of the mixture are significant to modelling the vorticity. I will now fit the model without the first two components (x_1 and x_2).

The alr transformed model: $U_i = \alpha + \beta_1 a_i^1 + \beta_2 a_i^2 + \varepsilon$

coefficients:

	Estimate	t value	$\Pr(>\|t\|)$
α	14.5099	24.357	$< 2 \times 10^{-16}$
β_1	1.4611	2.003	0.0487
β_2	1.2537	1.718	0.0897

The ilr transformed model: $U_i = \alpha + \beta_1 i_i^1 + \beta_2 i_i^2 + \varepsilon$

coefficients:

	Estimate	t value	$\Pr(>\|t\|)$
α	14.5099	24.357	$< 2 \times 10^{-16}$
β_1	-0.1467	-0.201	0.8412
β_2	-3.3250	-2.631	0.0103

These two models seem superior to the null model, each with an F-statistic of 3.482 on two and 78 degrees of freedom. This statistic produces a p-value of 0.03562 when testing if either of these models is superior to the null model. The full models, however, are superior to this reduced model as the residual sum of squares is significantly smaller, 1936.1 as opposed to 2242.1. This indicates that the first two components are related to the vorticity of a firework. This is the same conclusion that Aitchison (1986) drew from his analysis.

7.3 Example 3: Three mixture components with response

In this example, introduced by Cornell (1981, p. 95) a three-part composition and a binary process variable are used to model an observable non-negative response. The model suggested by Cornell (1981) is of the form $W = ax_1 + bx_2 + cx_3 + hzx_1 + izx_2 + jzx_3$. In the above model x_1, x_2, x_3 are the components of the composition, z is the process variable and a, b, c, h, i and j are the parameters estimated by the regression analysis. Cornell (1981) uses four design settings for the composition and uses two different process settings. The observed responses appear to be fictitious and generated by the mixture, and no comment is made about their origin by Cornell (1981).

Table 7.2: Table of mixture components, process variables and observed response for 8 designs Cornell (1981)

x_1	x_2	x_3	Process variable (z)	Observed response (W)
0.328	0.355	0.317	-1	14.3
0.482	0.201	0.317	-1	16.2
0.378	0.399	0.223	-1	12.6
0.532	0.245	0.223	-1	15.6
0.328	0.355	0.317	1	21.3
0.482	0.201	0.317	1	24.9
0.378	0.399	0.223	1	24.1
0.532	0.245	0.223	1	25.6

The composition will be denoted by X with x_1, x_2 and x_3 as the components. The data used can be found in Table 7.2.

Cornell found model $W = 27.1x_1 + 10.87x_2 + 16.32x_3 + 8.95zx_1 + 8.63zx_2 - 6.62zx_3 + \varepsilon$ (model 1) to fit the above data. This formula seems to contain redundancy, since the x_3 term can be rewritten as $1 - x_1 - x_2$. The method used does not explicitly take into account the constrained nature of the data. It produces an F-statistic of 54.2 on five and two degrees of freedom and a value of 1.405004 for the sum of residuals squared.

7.3.1 Compositional approach

The modern compositional approach is to first transform any compositions out of the simplex and into $\mathbb{R}$. We can do this via an alr or ilr transformation. Using the alr approach, with a_1 denoting the first component of the transform and a_2 the second, yields the following model (model 2):

$$W = \alpha + \beta_1 a_1 + \beta_2 a_2 + \beta_3 z + \beta_4 z a_1 + \beta_5 z a_2.$$

coefficients:

	Estimate	t value	$\Pr(>\|t\|)$
α	18.0041	34.052	0.000861
β_1	3.3010	3.417	0.076006
β_2	−2.4724	−3.155	0.087482
β_3	3.6892	6.978	0.019928
β_4	1.8800	1.946	0.191034
β_5	1.0929	1.395	0.297841

The corresponding ilr model, with i_1 denoting the first element of the transform and i_2 the second, yields the following model (model 3):

$$W = \alpha + \beta_1 i_1 + \beta_2 i_2 + \beta_3 z + \beta_4 z i_1 + \beta_5 z i_2.$$

coefficients:

	Estimate	t value	Pr(> |t|)
α	18.0041	34.052	0.000861
β_1	−4.0824	−4.413	0.047704
β_2	−1.0149	−0.705	0.554049
β_3	3.6892	6.978	0.019928
β_4	−0.5566	−0.602	0.608510
β_5	−3.6411	−2.528	0.127287

Models 2 and 3 both yield a p-value of 0.01704 on an F-statistic of 57.98 on five and two degrees of freedom. The sum of the residuals squared is 1.310635 for both of these models. These two models suggest that the interaction effects may not be very useful in modelling the data; therefore we model the effect of the process variable and the mixture of components on the observed response without any interaction terms. I will only include the alr-transformed model below as it will be easier to interpret. The model (model 4) is:

$$W = \alpha + \beta_1 a_1 + \beta_2 a_2 + \beta_3 z.$$

coefficients:

	Estimate	t value	Pr(> |t|)
α	18.0041	23.503	1.94×10^5
β_1	3.3010	2.359	0.07779
β_2	−2.4724	−2.178	0.09500
β_3	4.6500	11.214	0.00036

Model 4 has an F-statistic of 45.02 on three and four degrees of freedom. The p-value associated with a test of this model against the null model is 0.001537. The sum of the residuals squared for this model is 5.5024. The models created by this methodology appear to be better suited to predicting the response than the one given by Cornell (1981). The sum of the residuals

squared of models 2 and 3 (1.31) is less than that produced by model 1 (1.405), which indicates that the compositional models are better suited to represent the underlying data than the model of Cornell.

7.3.2 Reduced models

Cornell (1981) gives two separate models where the process variable is fixed at $z = 1$ or $z = -1$. The models are:

$$z = 1 : W = 36.06x_1 + 19.5x_2 + 9.7x_3$$

and

$$z = -1 : W = 18.16x_1 + 2.24x_2 + 22.94x_3.$$

No comment is made by Cornell (1981) on how suitable these models are, nor is any comment given on goodness of fit. The sums of the residuals squared for these two models are 1.102501 for the model with $z = 1$ and 0.3025035 for the model with $z = -1$

Using an alr transformation and fitting the data results in the following models. The models are in the form $W = \alpha + a_1\beta_1 + a_2\beta_2$ and $W = a + a_1b_1 + a_2b_2$ for $z = 1$ and $z = -1$ respectively, and are as follows.

coefficients:

	Estimate	t value	Pr($>\|t\|$)
α	21.6933	26.558	0.024
β_1	5.1811	3.472	0.179
β_1	−1.3795	−1.139	0.459

coefficients:

	Estimate	t value	Pr($>\|t\|$)
a	14.3149	21.317	0.0298
b_1	1.4210	1.158	0.4534
b_2	−3.5653	−3.582	0.1733

I will forego displaying the ilr transformations here as the results provide limited information. I will quote the sum of residuals squared in order to

compare these models to those models given by Cornell (1981). The sums of the residuals squared are 0.782057 and 0.5285776 for the alr-transformed models with $z = 1$ and $z = -1$ respectively.

Interestingly the model given by Cornell (1981) appears better suited to the $z = -1$ case than does the compositional model. The compositional model is better suited to the $z = 1$ case using a sum of residuals squared comparison. Overall, the model based on the compositional methods is better in this sense.

7.3.3 Further exploration using simulated data

This section will further compare the method proposed by Cornell (1981) with the method based on the additive log-ratio transformation. I will simulate data from fictitious models of each form and then compare the fit of each type of model to the data simulated. For ease of analysis I shall refer to as model C, a model of the form

$$W = ax_1 + bx_2 + cx_3 + hzx_1 + izx_2 + jzx_3,$$

as described by Cornell (1981). For the above model x_1, x_2, x_3 are the components of the composition, z is a process variable and a, b, c, h, i and j are the parameters estimated by the regression analysis. The model of the form

$$W = \alpha + \beta_1 a_1 + \beta_2 a_2 + \gamma_1 z + \gamma_2 z a_1 + \gamma_3 z a_2,$$

will be referred to as model A. In this model, (a_1, a_2) is the alr transformed composition, z is the process variable and $\alpha, \beta_1, \beta_2, \gamma_1, \gamma_2$ and γ_3 are the parameters of the model. Model C can be rewritten as:

$$\begin{aligned} W &= ax_1 + bx_2 + c(1 - x_1 - x_2) + hzx_1 + izx_2 + jz(1 - x_1 - x_2) \\ &= c + (a - c)x_1 + (b - c)x_2 + jz + (h - j)zx_1 + (i - j)zx_2. \end{aligned}$$

We can now compare the converted model C with model A. The difference between the models is just the use of x_1 and x_2 as covariates in model C and $a_1 = \log(x_1/x_3)$ and $a_2 = \log(x_2/x_3)$ as covariates in model A.

For the analysis I start by generating 150 random points from a uniform distribution with seed 04041992 in R using the `runif` command. These points are then split into triples, and 50 3-part compositions, X (with components x_1, x_2, x_3), are generated by applying the closure function to each triple. The 50 points are then repeated and a process value, z, of 1 is attached to the first 50 compositions and a process value, z, of -1 is attached to the next 50.

I will use model A and C to model two different simulated responses. The Model C response will be generated from the formula: $W = 10x_1 + 20x_2 + 30x_3 + 5zx_1 + 10zx_2 + 15zx_3 + \varepsilon$. The ε values are 100 random responses from a standard normal distribution generated in R, using the `rnorm` command with a seed of 19920404. The Model A response will be generated by: $10 + 10a_1 + 20a_2 + 5z + 5za_1 + 15za_2 + \varepsilon$. The alr transform of X is (a_1, a_2) and ε is as above, but with a seed value of 232323.

The R code used to simulate this data is as follows.

```
library(compositions)
set.seed(04041992)
table = c(runif(150))
comp=matrix(c(table,table),nrow=100,byrow=TRUE)
compC=acomp(comp)
p=c(rep(1,50),rep(-1,50))
proc=matrix(p,nrow=100,byrow=TRUE)
A = alr(compC)
set.seed(04041992)
responC = 10*compC[,1] + 20*compC[,2]+
30*compC[,3]+5*proc*compC[,1]+
10*proc*compC[,2]+15*proc*compC[,3]+ rnorm(100,0,1)
set.seed(232323)
responA =10+10*A[,1]+20*A[,2]+5*proc+5*proc*A[,1]+
15*proc*A[,2]+rnorm(100,0,1)
```

Model of type C

A model of form C produces the following:

$$W = 10.5602x_1 + 19.2161x_2 + 30.2201x_3 + 4.4077zx_1 + 10.3543zx_2 + 15.0283x_3 + \varepsilon.$$

This model has an F-statistic of 1932 on 5 and 94 degrees of freedom. The p-value associated with a test against the null model is sufficiently small to conclude that this model is a suitable one. The details of the model can be found below; note that the model is of the form $W = a + bx_1 + cx_2 + dz + fzx_1 + gzx_2$ which can easily be shown to be equivalent to model C.

coefficients:

	Estimate	t value	Pr($>\|t\|$)
a	30.2201	62.764	$< 2 \times 10^{-16}$
b	−19.6599	−24.943	$< 2 \times 10^{-16}$
c	−11.0040	−15.595	$< 2 \times 10^{-16}$
d	15.0283	31.212	$< 2 \times 10^{-16}$
f	−10.6206	−13.474	$< 2 \times 10^{-16}$
g	−4.6740	−6.624	2.16×10^{-9}

For this model all of the terms are significant, which can be expected from the simulated nature of the data.

A model of the form of model A produced the following:

$$W = 19.9097 - 1.9293a_1 + 0.1431a_2 + 9.9216z - 1.1441za_1 + 0.1868za_2 + \varepsilon.$$

This model has an F-statistic of 776 on five and 94 degrees of freedom. The F-statistic assures us that the full model is unsurprisingly better than the null model; the tests for whether each covariate of the model differs from 0 can be seen below.

coefficients:

	Estimate	t value	Pr($>\|t\|$)
α	19.9097	120.298	$< 2 \times 10^{-16}$
β_1	−1.9293	−12.759	$< 2 \times 10^{-16}$
β_2	0.1431	1.121	0.265
γ_1	9.9216	59.949	$< 2 \times 10^{-16}$
γ_2	−1.1441	−7.567	2.56×10^{-11}
γ_3	0.1868	1.463	0.147

This model suggests that $a_2 = \log(x_2/x_3)$ should not be included in the model and we could rerun the model with the second component removed. The above model indicates that removing a_2 and the interaction involving it could improve the appropriateness of the model. This is done by removing the x_2 from X and applying closure to the two remaining components.

A model of type A with x_2 removed to predict the response variable is now given. It is: $W = \alpha + \beta_1 a_1 + \gamma_1 z + \gamma_1 z a_1$, with estimates as below.

coefficients:

	Estimate	t value	$\Pr(>\|t\|)$
α	19.9290	120.199	$< 2 \times 10^{-16}$
β_1	-1.8394	-14.242	$< 2 \times 10^{-16}$
γ_1	9.9468	59.993	$< 2 \times 10^{-16}$
γ_2	-1.0269	-7.951	3.62×10^{-12}

The model has an F-statistic of 1274 on three and 96 degrees of freedom, which provides sufficient confidence that this model is superior to the null model.

Both the reduced form of model A and the full form of model C seem adequate, with the former being simpler.

Model of type A

A model of form C produces the following:

$$W = 52.987x_1 + 103.163x_2 - 129.268x_3 + 26.669x_1 + 72.979zx_2 - 87.776x_3 + \varepsilon.$$

This model has an F-statistic of 84.17 on five and 94 degrees of freedom. The F-statistic is sufficiently large to conclude that this model is superior to the null model. The details of the model can be found below; note that the model is of the form $W = a + bx_1 + cx_2 + dz + fzx_1 + gzx_2$ which can easily be shown to be equivalent to model C.

coefficients:

	Estimate	t value	$\Pr(>\|t\|)$
a	-129.268	-13.208	$< 2 \times 10^{-16}$
b	182.255	11.376	$< 2 \times 10^{-16}$
c	232.431	16.206	$< 2 \times 10^{-16}$
d	-87.776	-8.969	2.86×10^{-14}
f	114.445	7.143	1.91×10^{-10}
g	160.755	11.209	$< 2 \times 10^{-16}$

For this model all of the covariates appear to be significant.

A model of form A produced the following:

$W = 10.15687 + 9.99058a_1 + 20.04388a_2 + 4.83190z + 5.07889a_1 + 14.99780za_2 + \varepsilon.$

This model has an F-statistic of 38630 on five and 94 degrees of freedom. The F-statistic is large enough for us to be sure that this is superior to the null model.

coefficients:

	Estimate	t value	$\Pr(>\|t\|)$
α	10.15687	91.65	$< 2 \times 10^{-16}$
β_1	9.99058	98.67	$< 2 \times 10^{-16}$
β_2	20.04388	234.52	$< 2 \times 10^{-16}$
γ_1	4.83190	43.60	$< 2 \times 10^{-16}$
γ_2	5.07889	50.16	$< 2 \times 10^{-16}$
γ_3	14.99780	175.48	$< 2 \times 10^{-16}$

This model is practically a perfect fit, which is to be expected as the response was generated from a model of form A. Either of these models could be used to predict values for the simulated response, but the model of form A is more accurate.

It would appear that both could be used in a regression analysis. I will note that form A seems to predict a better model when the response is in form C, than a model of the form C manages for a response in the form of A.

Form A appears to be more robust under misspecification than form C. This section has shown that while both methodologies appear to work well, the models generated using compositional techniques are more robust than those generated ignoring restraints.

7.4 Example 4: A compositional response (national age distribution) potentially explained by past GDP and past unemployment

7.4.1 Introduction

I now provide an example where the response variable is compositional and the covariates are not compositional. I see whether the 2003 rate of unemployment and the 2003 Gross Domestic Product (GDP) per capita in

Purchasing Power Standards (PPS) can be used to help estimate the 2012 age composition of an EU population. This provides an example of the techniques dealt with in Section 6.3. The data used for the analysis were gathered from "`ec.europa.eu/eurostat`", and can be found in Table 7.3. PPS is a measurement of GDP which allows for the comparison of purchasing power among different areas during the same time period; it is an artificial currency unit which could purchase the same bundle of goods and services in different countries. I start by analysing the effects of GDP and unemployment rate on the proportions of different age groups for 30 European countries. I then analyse the effects of each one of GDP and unemployment rate on the proportions of age groups.

The age composition is the response variable and is denoted by Y, with the components labelled $Y^{(1)}$ for ages < 15, $Y^{(2)}$ for ages 15–65 and $Y^{(3)}$ for ages > 65. The unemployment rate and GDP per capita are the covariates and are denoted by P and Z respectively. I will start the analysis by plotting the response against the two covariates. The effect of unemployment rate and GDP size on the age structure of a population are shown in Figures 7.4 and 7.5 respectively.

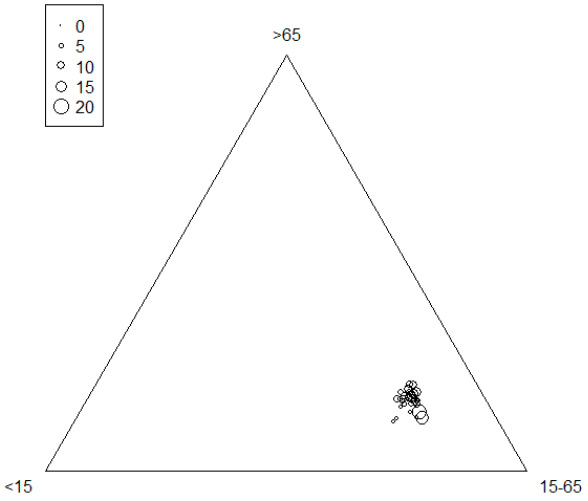


Figure 7.4: Ternary diagram showing the 2012 age composition scaled by the 2003 unemployment rates in different European countries.

Figures 7.4 and 7.5 are not very informative. Figure 7.4 seems to indicate (if anything) a population with a higher proportion of individuals in the 15–65 age category for higher unemployment rates.

Table 7.3: The unemployment rate and GDP per capita for 2003 and the proportion of individuals in different age groups in 2012 for 30 European countries.

Country	unemployment rate	GDP per capita (PPS)	age:<15	15-65	>65
Belgium	8.2	123	17	65.7	17.3
Bulgaria	13.7	33	13.4	67.8	18.8
Czech	7.8	77	14.7	69.1	16.2
Denmark	5.4	124	17.7	65	17.3
Germany	9.7	116	13.2	66.2	20.6
Estonia	10.3	52	15.5	66.8	17.7
Ireland	4.6	141	21.6	66.5	11.9
Greece	9.7	93	14.7	65.6	19.7
Spain	11.5	100	15.1	67.5	17.4
France	8.5	111	18.4	64.3	17.3
Croatia	14.2	56	15.1	67	17.9
Italy	8.4	112	14	65.2	20.8
Cyprus	4.1	94	16.5	70.7	12.8
Latvia	11.6	45	14.3	67.1	18.6
Lithuania	12.4	48	14.8	67.1	18.1
Luxembourg	3.8	240	17.1	68.9	14
Hungary	5.8	62	14.5	68.6	16.9
Malta	7.7	82	14.8	68.8	16.4
Netherlands	4.8	133	17.3	66.5	16.2
Austria	4.8	127	14.6	67.6	17.8
Poland	19.8	48	15.1	70.9	14
Portugal	7.4	78	14.9	66.1	19
Romania	7.7	31	15.8	68.1	16.1
Slovenia	6.7	83	14.3	68.9	16.8
Slovakia	17.7	55	15.4	71.8	12.8
Finland	9.0	114	16.5	65.4	18.1
Sweden	6.6	127	16.7	64.5	18.8
UK	5.0	123	17.6	65.6	16.8
Iceland	3.3	126	20.7	66.7	12.6
Norway	4.2	154	18.5	66.1	15.4

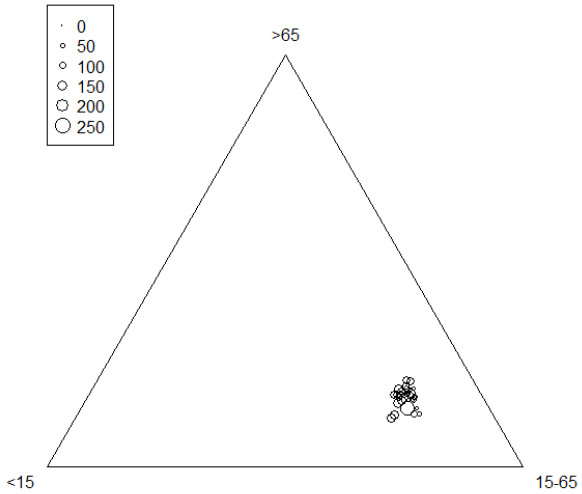


Figure 7.5: Ternary diagram showing the 2012 age composition scaled by the 2003 GDP size in different European countries.

The regression analysis can be performed via an alr or an ilr transformation. I will perform both of them.

7.4.2 Full model

The alr-transformed model, with the alr transformation of Y producing the vector (a_1, a_2), takes the form: $\text{alr}(Y) = \begin{pmatrix} a_1 \\ a_2 \end{pmatrix} = \alpha + P\beta_1 + Z\beta_2 + \varepsilon$.

coefficients:

	Estimate	t value	$\Pr(>\|t\|)$
α	-0.145328	-0.667	0.510
	1.3270041	8.212	8.11×10^{-9}
β_1	-0.007748	-0.573	0.571
	0.0014290	0.142	0.888
β_2	0.001697	1.361	0.185
	0.0005828	0.631	0.534

In each set of two estimates, the first refers to $a_1 = \log(Y^{(1)}/Y^{(3)})$ and the second to $a_2 = \log(Y^{(2)}/Y^{(3)})$.

An ilr transformed model, with an ilr transformation producing (s_1, s_2), is: $\text{ilr}(Y) = \begin{pmatrix} s_1 \\ s_2 \end{pmatrix} = \alpha + P\beta_1 + Z\beta_2 + \varepsilon$.

coefficients:

	Estimate	t value	$\Pr(>\|t\|)$
α	1.0410960	14.346	3.77×10^{-14}
	−0.4824172	−3.196	0.00353
β_1	0.0064894	1.441	0.1611
	0.0025799	0.275	0.78508
β_2	−0.0007877	−1.898	0.0685
	−0.0009307	−1.078	0.29053

The alr and ilr models transformed back into S^2 yield the following model:

$$\begin{pmatrix} Y^{(1)} \\ Y^{(2)} \\ Y^{(3)} \end{pmatrix} = \begin{pmatrix} 0.1534729 \\ 0.6690482 \\ 0.1774789 \end{pmatrix} \oplus P \odot \begin{pmatrix} 0.3314553 \\ 0.3345112 \\ 0.3340335 \end{pmatrix} \oplus Z \odot \begin{pmatrix} 0.3336457 \\ 0.3332742 \\ 0.3330800 \end{pmatrix} \oplus \epsilon. \quad \text{(A)}$$

It is important to remember that ϵ is a compositional Normal distribution derived by transforming a $N(0, I_2)$ distribution using an inverse alr or ilr transformation. The powering ($\odot$) operation is performed before the perturbation ($\oplus$) operation.

Here we use the `anova` command in R to assess the fit of the model. The null hypothesis for the test performed for each line of the anova test is: Given the proceeding $i-1$ variables, the ith covariate has no influence. As this test takes into account the order of the covariates it should be performed with each covariate in the last position. The results of the tests can be found below, first with P playing the role of the first added covariate and then with Z as the first added covariate:

ANOVA Table:

	Df	Pillai	approx F	num Df	den Df	$\Pr(>F)$
α	1	0.99757	5328.3	2	26	$< 2.2 \times 10^{-16}$
β_1	1	0.32761	6.3	2	26	0.005742
β_2	1	0.11769	1.7	2	26	0.196381
Residuals	27					

(Such a table applies to both the alr and ilr cases.)

ANOVA Table:

	Df	Pillai	approx F	num Df	den Df	Pr($> F$)
α	1	0.99757	5328.3	2	26	$< 2.2 \times 10^{-16}$
β_2	1	10.34561	6.9	2	26	0.004035
β_1	1	0.08465	1.2	2	26	0.316696
Residuals	27					

Neither covariate is significant when it is the last covariate added; when either unemployment rate or GDP is used to model the age composition, apparently the other is not needed. This suggests that we should attempt to model the age composition with unemployment as the only covariate, and then with GDP as the only covariate.

7.4.3 Unemployment rate as only covariate

I now attempt to model the age structure of the population in 2012 with the 2003 unemployment rate as the only covariate. I first provide the alr transformation and then the ilr one. The results can be found below. The alr model is: $\text{alr}(Y) = \begin{pmatrix} a_1 \\ a_2 \end{pmatrix} = \alpha + P\beta_1 + \varepsilon$.

coefficients:

	Estimate	t value	Pr($> \lvert t \rvert$)
α	0.12083	1.237	0.2264
	1.418423	20.092	$< 2 \times 10^{-16}$
β_1	-0.01974	-1.895	0.06858
	-0.002689	-0.357	0.724

The ilr model: $\text{ilr}(Y) = \begin{pmatrix} s_1 \\ s_2 \end{pmatrix} = \alpha + P\beta_1 + \varepsilon$

coefficients:

	Estimate	t value	$\Pr(>\|t\|)$
α	0.917537	27.384	$< 2 \times 10^{-16}$
	-0.628398	-9.400	3.7×10^{-10}
β_1	0.012056	3.374	0.00218
	0.009156	1.284	0.21

The form of the model after transformation into S^2 is as follows:

$$\begin{pmatrix} Y^{(1)} \\ Y^{(2)} \\ Y^{(3)} \end{pmatrix} = \begin{pmatrix} 0.1802887 \\ 0.6599423 \\ 0.1597690 \end{pmatrix} \oplus P \odot \begin{pmatrix} 0.3292582 \\ 0.3349199 \\ 0.3358219 \end{pmatrix} \oplus \epsilon. \tag{B}$$

In order to determine whether this model is suitable the anova command in R will once again be used. The table produced can be found below:

ANOVA Table:

	Df	Pillai	approx F	num Df	den Df	$\Pr(>F)$
α	1	0.99742	5210.1	2	27	$< 2.2 \times 10^{-16}$
β_1	1	0.30225	5.8	2	27	0.007762
Residuals	28					

The table shows us that we can be reasonably sure that β_1 is different from 0, and thus conclude that the 2012 population composition is dependent on the 2003 unemployment rates.

7.4.4 GDP as only covariate

I will now model the age structure in 2012 by the 2003 GDP per capita, again exploring both an alr transformation and an ilr transformation. The

alr model is: $\text{alr}(Y) = \begin{pmatrix} a_1 \\ a_2 \end{pmatrix} = \alpha + Z\beta_2 + \varepsilon.$

coefficients:

	Estimate	t value	Pr($>\|t\|$)
α	-0.2561134	-2.580	0.0154
	1.3474355	18.410	$< \times 10^{-16}$
β_2	0.0021619	2.314	0.0283
	0.0004971	0.721	0.477

The ilr model is: $\text{ilr}(Y) = \begin{pmatrix} s_1 \\ s_2 \end{pmatrix} = \alpha + Z\beta_2 + \varepsilon$

coefficients:

	Estimate	t value	Pr($>\|t\|$)
α	1.133880	33.255	$< 2 \times 10^{-16}$
	-0.4455304	-6.511	4.68×10^{-07}
β_2	-0.001177	-3.667	0.00102
	-0.0010855	-1.685	0.103

The resulting model in S^2 for both the alr and ilr transformation is as follows:

$$\begin{pmatrix} Y^{(1)} \\ Y^{(2)} \\ Y^{(3)} \end{pmatrix} = \begin{pmatrix} 0.1376929 \\ 0.6844218 \\ 0.1778853 \end{pmatrix} \oplus Z \odot \begin{pmatrix} 0.3337587 \\ 0.3332035 \\ 0.3330379 \end{pmatrix} \oplus \epsilon. \qquad \text{(C)}$$

In order to determine whether this model is suitable the anova procedure in R will be used. The table produced can be found below:

ANOVA Table:

	Df	Pillai	approx F	num Df	den Df	Pr($>F$)
α	1	0.99737	5123.0	2	27	$< 2.2 \times 10^{-16}$
β_2	1	0.32743	6.6	2	27	0.004726
Residuals	28					

The Anova table suggests that β_2 is non-zero; thus we would conclude that the population composition is dependent on the GDP rate.

A feature of the models (A), (B) and (C) is that the compositions powered by P or Z are always approximately $(1/3, 1/3, 1/3)$. If these compositions had been exactly $(1/3, 1/3, 1/3)$, the covariates P and Z would have had no effect on the response. As it is, they have very little effect on the response.

7.5 Discussion

In this dissertation I have discussed certain aspects of the statistical analysis of compositions. I started by defining the term 'compositional data' and giving a brief history of compositional data analysis. I then gave two methods of visualising compositional data. The vector space of compositional data, known as the simplex, and its associated vector space operations, perturbation and powering, were then defined.

Next the three main transformations from the simplex to unconstrained real space were described. These transformations each have their own advantages and disadvantages. In Chapter 4 several further properties of compositions were described and the simplex further explored. Chapter 5 demonstrated how hypothesis testing could be performed with compositional data, and introduced a 'lattice of hypotheses' methodology to be followed in tests of hypothesis.

With the information presented in the previous chapters I was then able to describe the expansion of regression analysis to compositional data. The methodology proved to be quite simple, with the main steps being a transformation of the compositional data into unconstrained space and the use of familiar methods of analysis within this space.

In the final chapter, I demonstrated several examples of the application of regression involving compositional data. I performed regression analysis using standard regression techniques with a composition transformed via the alr or an ilr transformation. The chapter started with a simple example where the composition was the covariate and the response was non-compositional.

The next application also involved a composition as a covariate and revisited an example of Aitchison (1986), but using both alr and ilr methods. It highlighted the fact that the alr results are more interpretable than those given by an ilr transformation. An example of Cornell (1981) was then examined and his 'non-compositional' method of analysing a mixture was compared with compositional techniques. Both methods produced quite accurate results on simulated data, but the methods based on compositional analysis proved to be slightly superior under misspecification.

In the final example, a composition was the response and the covariates were non-compositional. Here the alr and ilr transformations produce the

same model when transformed back into the simplex.

This dissertation has aimed to convey an understanding of what a composition is and explain how to allow for the compositional nature of the data. It has dealt with the transformation of components and explained how regular techniques can then be applied to transformed compositional data. Finally it has shown several examples of applications, in order to demonstrate how the transformations are performed, and how regression analysis can be performed on compositional data.

References

Aitchison, J. (1982). The statistical analysis of compositional data. *Journal of the Royal Statistical Society Series B (Methodological)*, **44**(2), 139–177.

Aitchison, J. (1986). *The Statistical Analysis of Compositional Data*. Chapman and Hall, London.

Aitchison, J. (2003). A concise guide to compositional data analysis. Accessed[16/09/2015] [http://ima.udg.edu/activitats/codawork05/A_concise_guide_to_compositional_data_analysis.pdf].

Aitchison, J. (2005). Compositional data analysis: Where are we and where should we be heading? Accessed[16/09/2015] [http://dugi-doc.udg.edu:8080/bitstream/handle/10256/647/Aitchison.pdf?sequence=1].

Aitchison, J. (2008). The single principle of compositional data analysis, continuing fallacies, confusions and misunderstandings and some suggested remedies. Online. Accessed[16/09/2015] [http://dugi-doc.udg.edu:8080/handle/10256/706].

Aitchison, J. and Brown, J. (1969). *The Lognormal Distribution with References to its Uses in Economics*. Cambridge University Press, Cambridge.

Aitchison, J. and Greenacre, M. (2002). Biplots of compositional data. *Journal of the Royal Statistical Society Series C*, **51**(4), 375–392.

Aitchison, J. and Ng, K. (2005). The role of peturbation in compositional data analysis. *Statistical Modelling*, **5**(2), 173–185.

Aitchison, J. and Shen, S. (1980). Logistic-normal distribution. some properties and uses. *Biometrika*, **67**(2), 261–272.

Aitchison, J., Barcelo-Vidal, C., Mateu-Figueras, G., and Pawlowsky-Glahn, V. (2000). Logratio analysis and compositional distance. *Mathematical Geology*, **32**(3), 271–275.

Bacon-Shone, K. (2011). A short history of compositional data analysis. In V. Pawlowsky-Glahn and A. Buccianti, editors, *Compositional Data Anaysis: Theory and Applications*. John Wiley & Sons, Chichester, UK.

Chayes, F. (1960). On correlation between variables of constant sum. *Journal of Geophysical Research*, **65**(12), 4185–4193.

Chayes, F. (1962). Numerical correlation and petrographic variation. *The Journal of Geology*, **70**(4), 440–452.

Cornell, J. (1981). *Experiments with Mixtures*. John Wiley & Sons, New York.

Cox, D. and Snell, E. (1989). *Analysis of Binary Data*. Chapman and Hall, London, 2nd edition.

Dumuid, D., Stanford, T. E., Martin-Fernandez, J.-A., Pedisic, Z., Maher, C. A., Lewis, L. K., Hron, K., Katzmarzyk, P. T., Chaput, J.-P., Fogelholm, M., Hu, G., Lambert, E. V., Maia, J., Sarmiento, O. L., Standage, M., Barreira, T. V., Broyles, S. T., Tudor-Locke, C., Tremblay, M. S., and Olds, T. (2017). Compositional data analysis for physical activity, sedentary time and sleep research. *Statistical Methods in Medical Research*, **Article 0962280217710835**.

Egozcue, J., Pawlowsky-Glahn, V., Mateu-Figueras, G., and Barcelo-Vidal, C. (2003). Isometric logratio transformations for compositional data analysis. *Mathematical Geology*, **35**(3), 279–300.

Hron, K., Filzmoser, P., and Thompson, K. (2012). Linear regression with compositional explanatory variables. *Journal of Applied Statistics*, **39**(5), 1115–1128.

Kotz, S., Balakrishnan, N., and Johnson, N. (2000). *Continuous Multivariate Distributions. Volume 1, Models and Applications*. John Wiley & Sons, New York.

Lea, M. and Leite, C. (2016). Applying compositional data methodology to nutritional epidemiology. *Statistical Methods in Medical Research*, **25**, 3057–3065.

McAlister, D. (1897). The law of geometric mean. *Proceedings of the Royal Society of London*, **29**, 625–645.

Pawlowsky-Glahn, V., Egozcue, J. J., and Tolosana-Delgado, R. (2007). Lecture notes on compositional data analysis. Accessed[16/09/2015] [http://dugi-doc.udg.edu/bitstream/handle/10256/297/CoDa-book.pdf?sequence=1].

Pawlowsky-Glahn, V., Egozcue, J., and Tolosana-Delgado, R. (2015). *Modeling and Analysis of Compositional Data*. John Wiley & Sons, New York.

Pearson, K. (1897). On a form of spurious correlation which may arise when indices are used in the measurement of organs. *Proceedings of the Royal Society of London*, **LX**, 489–502.

Simpson, S. J., Batley, R., and Raubenheimer, D. (2003). Geometric analysis of macronutrient intake in humans: the power of protein? *Appetite*, **41**(2), 123–140.

Tanner, J. (1949). Fallacy of per-weight and per-surface area standards, and their relation to spurious correlation. *Journal of Applied Physiology*, **2**(1), 1–15.

Van den Boogaart, K. G. and Tolosana-Delgado, R. (2013). *Analyzing Compositional Data with R*. Springer, Berlin.