Comparing Multivariate Regression Methods For Compositional Data

Through Simulation Studies & Applications

Christoffer Långström

Examensarbete, 15hp, VT 2017
Kandidatexamen i matematik, 180hp
Institutionen för matematik och matematisk statistik
Summary

Compositional data, where measurements are vectors with each component constituting a percentage of a whole, is abundant throughout many disciplines of science. Consequently, there is a strong need to establish valid statistical procedures for this type of data. In this work the basic theory of the compositional sample space is presented and through simulation studies and a case study on data from industrial applications, the current available methods for regression as applied to compositional data are evaluated. The main focus of this work is to establish linear regression in a way compatible with compositional data sets and compare this approach with the alternative of applying standard multivariate regression methods on raw compositional data. It is found that for several data sets, the difference between ‘naive’ multivariate linear regression and compositional linear regression is negligible; while for others (in particular where the dependence of covariates is not strictly linear) the compositional regression methods are shown to be stronger.

Sammanfattning

Kompositionsdata, där mätningar är vektorer med varje komponent en procentenhet av någon helhet, förekommer inom många naturvetenskapliga forskningsområden. Därför finns det ett stark behov att etablera statistiska metoder som är lämpliga för denna typ av data. I det här arbetet presenteras grundläggande teori för det kompositionella urvalsrummet; vidare utvärderas nuvarande metoder genom simuleringstudier och tillämpningar från industri. Fokus ligger på att etablera linjär regression på kompositionsdata och att jämföra dessa metoder med att nyttja traditionell linjär regression. Slutsatsen dras att för flera typer av data är skillnaderna mellan metoderna försumbar, men för andra typer (framförallt där beroendet på förklarande variabler inte är strikt linjärt) är kompositionella regressionsmetoder starkare.
Contents

1 Introduction 6

2 Theoretical Preliminaries 8
   2.1 Multivariate Regression in \( \mathbb{R}^D \) ................................. 8
   2.2 Multivariate Linear Models ............................................ 8
   2.3 The Simplex Sample Space ............................................. 10
   2.4 Statistics In The Simplex Space ..................................... 11
      2.4.1 The Principles Of Simplical Analysis ............................ 11
   2.5 Transforms From The Simplex to Euclidian Space .................. 12
      2.5.1 Handling data with zero values ................................. 14
   2.6 Linear Regression In The Simplex ................................... 14
      2.6.1 Transformed Regression & The Logistic Normal Distribution .. 14
   2.7 Confidence Intervals, Resampling, & Bootstrap Methods ............ 16

3 Methods 19
   3.1 Approach To Applying Linear Regression In The Simplex .......... 19
   3.2 Simulations Of Compositional Data .................................. 20
      3.2.1 Types Of Simulated Distributions ............................... 20
      3.2.2 Simulated Distributions ........................................... 20
   3.3 Case Study: Industrial Processing .................................... 22

4 Results 25
   4.1 Simulation Study ....................................................... 25
      4.1.1 Simulation 1: 3-Component Multivariate Normal Distribution .. 25
      4.1.2 Simulation 2: 3-Component Univariate Uniform Distribution ... 27
   4.2 Results of Case Study .................................................. 30

5 Discussion & Conclusions 34
   5.1 Simulation Study ....................................................... 34
   5.2 Industrial Data ......................................................... 34

6 Acknowledgements 36

7 Appendix 37
   7.0.1 Proof of Proposition 1, section 2.1 .............................. 37
1 Introduction

In many sections of natural science, multivariate data from experiments are recorded in proportions, where components are the percentages of some whole. An example of this is in geology, where the absolute weight of a gathered rock sample may not be interesting to the researcher, only the mineral compositions of the sample is relevant. Mathematically, this corresponds to restricting the sample space to the unit simplex, or some scalar multiple thereof. This inevitably leads to some questions regarding what the statistical methodology should be when the data is restricted to such a space. A common conception of practitioners of statistics unfamiliar with the theory of compositional analysis is that the restriction of the sample space is irrelevant, and that standard multivariate methods can be applied without second thought. In this work we will, through a combination of simulation studies and applications to data from applications, investigate both the practical and theoretical validity of such statements and to what extent the standard methodology remains valid in the context of the simplex. The main source of information on this topic has been [1].

In its nature, compositional analysis comprises a subset of multivariate statistics where each measured data vector of size \( D \) is confined to the unit \( D - 1 \) simplex. Data confined to this set will inevitably cause the compromise of certain statistical methods of analysis. Throughout the 20th century questions as to what extent this restricted data will violate methods of analysis were raised, with Karl Pearson [4] being one of the critics. A firm theoretical foundation for the subject was laid out in several publications by Aitchison, see [2], [3], (compiled and expanded upon in [1]) where several approaches to compositional analysis were suggested which could be broadly divided into two categories. One approach was to make use of one-to-one transforms, mapping the unit simplex \( S^D \) to the full space of \( \mathbb{R}^{D-1} \), where familiar multivariate methods would be applicable to the transformed data, and any conclusion drawn from this data could then be taken back to the simplex. A second approach was the idea to remain in the simplex and instead develop the familiar statistical concepts in terms of operations that would preserve the structure of the simplex. The equivalence of these two methods in terms of inference made was then proved, which made the choice of approach selected by the statistician largely a matter of preference. We now consider a typical example of when compositional data occurs in practice.

Example 1: A geologist has gathered rock specimens of a certain type, each sample consisting of certain quantities of mineral compounds. A single rock specimen is made up of quartz, clay minerals, iron oxides and various compounds classified as "other". Table 1 shows the mineral composition of the samples gathered.

<table>
<thead>
<tr>
<th>Quartz</th>
<th>Clay</th>
<th>Iron Ox.</th>
<th>Other</th>
</tr>
</thead>
<tbody>
<tr>
<td>0.4500</td>
<td>0.2206</td>
<td>0.1206</td>
<td>0.2088</td>
</tr>
<tr>
<td>0.4940</td>
<td>0.2141</td>
<td>0.0921</td>
<td>0.1998</td>
</tr>
<tr>
<td>0.4763</td>
<td>0.2167</td>
<td>0.1035</td>
<td>0.2034</td>
</tr>
<tr>
<td>0.4023</td>
<td>0.2276</td>
<td>0.1514</td>
<td>0.2186</td>
</tr>
<tr>
<td>0.4247</td>
<td>0.2243</td>
<td>0.1370</td>
<td>0.2140</td>
</tr>
<tr>
<td>0.4496</td>
<td>0.2206</td>
<td>0.1208</td>
<td>0.2089</td>
</tr>
<tr>
<td>0.3777</td>
<td>0.2312</td>
<td>0.1673</td>
<td>0.2237</td>
</tr>
<tr>
<td>0.4066</td>
<td>0.2270</td>
<td>0.1486</td>
<td>0.2177</td>
</tr>
<tr>
<td>0.4119</td>
<td>0.2262</td>
<td>0.1452</td>
<td>0.2167</td>
</tr>
<tr>
<td>0.3700</td>
<td>0.2324</td>
<td>0.1724</td>
<td>0.2253</td>
</tr>
</tbody>
</table>

Some questions that are of immediate interest are:

- Is there such a thing as an 'average' composition of the rock types?
- Can we describe the variation around this average of the compositions?
If another rock sample is gathered from a different location, can we say with a certain degree of confidence that the new sample is of the same rock type as those previously gathered?

The topological considerations of the matter would suggest that since \( S^D \) is a subset of \( \mathbb{R}^D \), any statistical method of analysis applicable in \( \mathbb{R}^D \) would then be equally valid when the data is restricted to \( S^D \). An early indication that this was not the case was the so called negative bias problem. As a consequence of the definition of \( S^D \), in particular the unit sum constraint of the components, we have that for \( \mathbf{x} = (x_1, x_2, ..., x_D) \in S^D \) that
\[
\text{cov}(x_1, x_1 + x_2 + ... + x_d) = \text{cov}(x_1, 1) = 0
\]
which implies that
\[
\text{cov}(x_1, x_2) + \text{cov}(x_1, x_3) + ... \text{cov}(x_1, x_D) = -\text{var}(x_1).
\]
Excluding the trivial case where the right hand side is zero, it is strictly negative, which forces at least one of the covariances on the right hand side to be negative. Thus the range of \( \text{cov}(x_i, x_j) \) is not \([-1, 1]\) as one would be accustomed, and problem of interpretation of any statistical inference immediately arises. The following example shows a particularly awkward case.

Example 2: Consider a two-part composition \( \mathbf{x} = (x_1, x_2) \). Then
\[
\text{cov}(x_1, x_2) = -\text{var}(x_1)
\]
and
\[
\text{cov}(x_2, x_1) = -\text{var}(x_2)
\]
so \( \text{var}(x_1) = \text{var}(x_2) \) and thus
\[
\text{corr}(x_1, x_2) = -1
\]
so the correlation is constant, and does not have the standard range of \([-1, 1]\).

It is clear from these results and the fact the inter-component covariances are negatively biased, that statistical independence between components is not trivially defined. A number of attempts to define compositional independence are laid out in [1, chapter 3, p. 64]. In this study two major sources of data are considered. In 3.2.1 simulations are used to generate data of various forms to test. We look at a case study of industrial data from current research, where industrial processing has produced distributions of rock samples that are then estimated via laser topography scanning, which can then be used to make inferences about the original distribution.
2 Theoretical Preliminaries

2.1 Multivariate Regression in $\mathbb{R}^D$

Multivariate regression is concerned with modelling a vector of dependent variables, $Y$ as a function of a vector of independent variables $X$ called covariates. The model is then relating the dependent variables to the independent variables, via an unknown parameter matrix $\beta$, that is

$$ Y = f(X, \beta) + E $$

with the constraint that $E(Y|X) = f(X, \beta)$. Solutions to this system depends heavily on the nature of the function $f$, and what numerical methods are available for such functions. If the function is known, then appropriate methods can be chosen, otherwise it must be guessed from the available observations of $Y$ and $X$. The choice of function $f$ will be the deciding factor in the quality of our model. In estimating the coefficient matrix $\beta$ in (1), the sample size will be crucial. If the number of measurements $n$ is low in relation to the number of parameters to be estimated, then we may encounter overfitting, in that an estimate of $\beta$ can be found but the model will have poor predictive performance; essentially the model will fit the training data well but cannot accurately predict behaviour in unobserved data.

In general, there are some assumptions made on the observations of $Y$ and $X$. First, it is implicitly assumed that the sample is representative of the overall population, in order to motivate analysis of the data set. Secondly, the deviation $E$ of $Y$ from $f(X, \beta)$ is a homoscedastic (constant variance) random variable with zero mean, referred to as the error. In the case that the function $f$ in (1) is linear, there is a rich framework of theory and models to perform analysis. This will be the main topic of this work.

2.2 Multivariate Linear Models

The standard method of multivariate multiple linear regression is a natural extension of multivariate linear regression, which models a single dependent variable as a linear function of multiple covariates, to the case where we have a vector of dependent variables. No restriction is made on the data, other than for it to be in real space. Formally, the problem is to model $k$ dependencies

$$ Y = (Y_1, Y_2, ..., Y_k) $$

as a function, linear in its coefficients, of covariates

$$ X = (X_1, X_2, ..., X_d) $$

with the addition of a vector of errors $E = (\epsilon_1, \epsilon_2, ..., \epsilon_k)$ with zero mean and covariance matrix $\Sigma$, which are (unobserved) random variables of noise. The model can be formulated in a familiar way as a system of equations

$$ Y_1 = \beta_{11} + \beta_{12}X_1 + \cdots + \beta_{1k}X_d + \epsilon_1 $$
$$ Y_2 = \beta_{21} + \beta_{22}X_1 + \cdots + \beta_{2k}X_d + \epsilon_2 $$
$$ \vdots $$
$$ Y_k = \beta_{k1} + \beta_{k2}X_1 + \cdots + \beta_{kd}X_d + \epsilon_k $$

or in concise matrix form

$$ Y = \beta X + E $$

with $\beta = \{\beta_{ij}\}, 1 \leq i \leq k, 1 \leq j \leq d$. The standard approach to finding an estimate for $\beta$ is through the least squares, in which a hyperplane in $\mathbb{R}^K$ is sought such that sum of squared residuals between observed values and fitted modelsled values is minimized. This is known as the ordinary least squares (OLS) estimate of $\beta$, and under certain assumptions on the distribution of the error in (2) this estimate coincides with the maximum likelihood estimates, namely the assumption that the errors are distributed as $\mathcal{N}(0, \Sigma)$. If so, these estimates of $\beta$ and $\Sigma$ can be easily derived.
Proposition 2.2.1. The maximum likelihood estimate $\hat{\beta}$ of $\beta$ under the model described in (2) is

$$\hat{\beta} = (X^TX)^{-1}X^T$$

(3)

Once a model has been fitted as $\hat{Y} = \hat{\beta}X$ the residuals are given by

$$\hat{E} = (I - X(X^TX)^{-1}X^T)Y$$

For a sample of size $N$, an unbiased estimator of $\Sigma$ is given by

$$\hat{\Sigma} = \frac{1}{N}\hat{E}^T\hat{E} = \frac{1}{N}(Y - X\beta)^T(Y - X\beta)$$

Proofs of these assertions can be found in the appendix.

In order for (3) to have a well defined solution, it is necessary for the matrix product $X^TX$ to be invertible. This has two important implications.

1. The data gathered should be taken for sufficiently many distinct values of the covariates.

2. Each covariate $X_i \in X$ should not be expressible as a linear combination of any of the other covariates.

Most notably, condition 2 will play an important role in the study of compositional data, wherein data is linearly dependent by definition.

It is worth noting that the dependency relation is not necessarily linear in its covariates, only in the coefficients $\beta_{ij}$. For example, the following model would also be acceptable

$$Y_i = \beta_{i1} + \beta_{i2}\log(X_1) + \beta_{i3}X_3^2 + \ldots + \epsilon_i$$

and in fact we will consider such models in our investigation of compositional distributions.
Formally we define the unit simplex $S^D$ as the subset of $\mathbb{R}^D$ such that

$$S^D = \{(x_1, x_2, \ldots, x_D): \sum_{i=0}^{D} x_i = 1, x_i \geq 0 \ \forall i\}.$$ 

We begin by considering properties we want to develop in this sample space. Wishing to mirror the familiar operations of scalar multiplication and vector addition in euclidian space, Aitchison [1] provided the following analogies to $S^D$:

**Definition 1.** Let $x, y \in S^D$. Then the permutation of $x$ with $y$ is defined

$$x \oplus y = \left(\frac{1}{\sum_{i=0}^{D} x_i y_i} \prod_{i=0}^{D} x_i y_i\right)$$

Let $b$ be a scalar. The power of $x$ with $b$ is defined

$$x \otimes b = \left(\frac{1}{\sum_{i=0}^{D} x_i^b} \prod_{i=0}^{D} x_i^b\right)$$

It is possible to show that $\langle S^D, \oplus \rangle$ forms an Abelian group and together the operations define a vector space $\langle S^D, \oplus, \otimes \rangle$. One can also define an inner product in this space, along with corresponding notions of norm and metric to form a complete vector space.

Notice that any vector $v = (v_1, v_2, \ldots, v_d) \in \mathbb{R}^d$ can be considered in terms of relative size of components by dividing through with the sum $v_1 + v_2 + \ldots + v_d$. This transformation from $\mathbb{R}^D$ to $S^D$ by dividing through with the sum of the components occurs frequently, so for notational simplicity it is useful to define the following operation.
Definition 2. Let $x$ be a positive real vector with components $x_1, x_2, ..., x_D$. The closure operator $C(\cdot)$ is defined as

$$C(x) = \frac{x}{\sum_{i=0}^{D} x_i}$$

2.4 Statistics In The Simplex Space

2.4.1 The Principles Of Simplical Analysis

Here we develop basic statistical properties and extend familiar notions to data in $\mathbb{S}^D$. In developing statistical tools for the simplex, we must address some core concepts vital to this sample space. First is the idea of scale invariance. Given two samples $x = (x_1, x_2, ..., x_D)$ and $y = (y_1, y_2, ..., y_D)$, if

$$C(x_1, x_2, ..., x_D) = C(y_1, y_2, ..., y_D)$$

then in the context of compositional analysis $x = y$, that is they are compositionally equivalent, denoted $x \sim y$. This is known as the principle of scale invariance, which formalizes the idea that the absolute quantity of any component in a sample is irrelevant to composition, only its proportion to the whole is of interest. Further, in many applications the user may be interested in only parts of the original composition, say $(x_1, x_2, x_3) \in (x_1, x_2, ..., x_D)$. The interest of the practitioner then is that any statistical inference made on the subset, should match inference made on the original set when viewed in terms of the components $x_1, x_2, x_3$. Note that taking $C(x_1, x_2, x_3)$ results in a projection from $\mathbb{S}^D$ to $\mathbb{S}^3$. Mathematically, we require that any meaningful function $f$ of a compositional data set must satisfy $f(W) = f(w)$ whenever $W \sim w$, or equivalently

$$f(pw) = f(w), \quad \text{for every } p > 0.$$  

To be precise, it is required that $f$ must be invariant under the group of scalar transformation. An important example of such a family of functions are component ratio functions, such as $H(x) = (x_1, x_2, ..., x_{d-1})/x_d$

This fact will be of great use in section 2.5 were we discuss transforms of compositions. The unit sum constraint inherent to the simplex requires an alternate formulation of the mean of a compositional random variable. Using the metrics defined in 2.3 we can develop an analogous concept in the form of a center of a random composition.

Definition 3. The center of a compositional random variable $X$ in $\mathbb{S}^D$ is defined as

$$\text{Cen}(X) = C(\exp(E(\log(X))))$$

We now give an unbiased estimator of the center as given by [7] which can be intuitively described as the closure of the component wise geometric mean.

Proposition 2.4.1. Let $X \in \mathbb{S}^D$ with components $x_1, x_2, ..., x_D$, and $x_1, x_2, ..., x_n$ be a random sample of $X$. Then

$$\bar{X} = \frac{1}{n} \otimes \bigoplus_{i=1}^{n} x_i = C \left( \left( \prod_{i=1}^{n} x_{i1} \right)^{\frac{1}{n}}, \left( \prod_{i=1}^{n} x_{i2} \right)^{\frac{1}{n}}, ..., \left( \prod_{i=1}^{n} x_{iD} \right)^{\frac{1}{n}} \right)^T$$

is an unbiased estimator of $\text{Cen}(X)$.

Defining the variance of a compositional random variable is more complicated, with several definitions suggested in [1], and [5]. In this work we will utilize the total covariance structure as defined in [1]:

11
Definition 4. The total covariance matrix for a compositional variable \( X = (X_1, X_2, ..., X_D) \) is defined as

\[
T = \{\tau_{ij}\}, \quad i, j \in \{1, 2, 3, ..., D\}
\]

where

\[
\tau_{ij} = \text{Var} \left( \log \left( \frac{X_i}{X_j} \right) \right)
\]

The sample estimate of \( T \) is obtained by replacing the variance in the second expression with the corresponding sample estimate. Basic properties of the structure of \( T \) is that it is a symmetric matrix with a zero diagonal. The symmetry comes from the fact that

\[
\text{Var} \left( \log \left( \frac{X_i}{X_j} \right) \right) = \text{Var} \left( \log(X_i) - \log(X_j) \right) = \text{Var} \left( \log(X_j) - \log(X_i) \right) = \text{Var} \left( \log \left( \frac{X_j}{X_i} \right) \right).
\]

The matrix \( T \) provides information about the relative variance between the components of \( X \), in that lower values \( \tau_{ij} \) indicates that changes in component \( i \) does not reflect in a substantial change in component \( j \), and vice versa. Due to the symmetric nature of the matrix, it has been suggested by Aitchison in [1] to use the lower half of the matrix to display the expected value of the quotient logarithms of components, that is letting

\[
\xi_{ij} = E \left( \log \left( \frac{X_i}{X_j} \right) \right)
\]

and defining an extension of \( T \) as such

\[
\Gamma = \begin{pmatrix}
0 & \tau_{1,2} & \cdots & \tau_{1,D} \\
\xi_{1,2} & 0 & \cdots & \tau_{2,D} \\
\vdots & \vdots & \ddots & \vdots \\
\xi_{1,D} & \xi_{2,D} & \cdots & 0
\end{pmatrix}
\]

The purpose of this extension is to provide information on the average relative shift in the relation of components, with the placement of the \( \xi_{ij} \) made so that each mean corresponds to the mirrored variance in the array. If \( \xi_{ij} < 0 \) then this would indicate that the \( i, j \) quotient is less than one on average, making the logarithm negative. This would then in turn indicate that component \( i \) is overall smaller than component \( j \), a claim which can then be further corroborated by comparing the value of \( \xi_{ij} \) with the corresponding value of \( \sqrt{\tau_{ij}} \). If \( \xi_{ij} \) is significantly larger than \( \sqrt{\tau_{ij}} \), then this would indicate that a component \( i \) is indeed overall 'smaller' than component \( j \). In light of this, it makes sense to use the relative standard deviations when displaying the matrix. For this purpose, we define \( \Gamma \) to be the matrix that is obtained from the above extension of \( T \) with the variances replaced by the standard deviation. It is this variant we will be using in our analysis in future sections.

2.5 Transforms From The Simplex to Euclidian Space

These transforms are designed with the idea of taking data obtained from \( S^D \) into the real euclidian space, in order to legitimize the use of standard multivariate analysis methods of the data. By choosing a one-to-one transform we can ensure that the transform is invertible and any inference done in the euclidian space can be suitably expressed in terms of the simplex.

The unit sum constraint of components in \( S^D \) dictates that any one component can be found if the other \( D - 1 \) are known. This implies that any set of components can at most span a space of dimension \( D - 1 \), and so it is natural to consider transformations from \( S^D \) to \( \mathbb{R}^{D-1} \), with an inverse transform using the unit sum constraint to reconstruct the missing component, although other transformations that preserve the \( D \) number of components are also considered. This immediately leads to the question: which component should be chosen to be be excluded and how will this affect the analysis. Fortunately, as described in [1], the choice of divisor is arbitrary, as each choice provides identical inference. This equivalence is demonstrated in [1, ch. 4.5.6, p. 234].
The simplest and most easily implemented transformation is the *additive log ratio transform* which this work will be largely centred around.

**Definition 5.** Let \( \mathbf{x} = (x_1, x_2, \ldots, x_D) \in \mathbb{S}^D \). The transformation \( \text{alr} : \mathbb{S}^D \to \mathbb{R}^{D-1} \) is defined as

\[
y = \text{alr}(\mathbf{x}) = \left( \log \left( \frac{x_1}{x_D} \right), \log \left( \frac{x_2}{x_D} \right), \ldots, \log \left( \frac{x_D-1}{x_D} \right) \right)
\]

where \( x_D \) is the last element of \( \mathbf{x} \). The inverse transformation is

\[
\text{alr}^{-1}(y) = C(e^{y_1}, e^{y_2}, \ldots, e^{y_{D-1}}, 1)
\]

The difference in choosing an element other than the last one for division is that the number 1 in the inverse transformation vector occupies the same position as the chosen divisor did before the transform. The relative simplicity of the \( \text{alr} \) transform makes it appealing as an obvious candidate for quick implementation and analysis. The most immediate drawback is that it requires that at least one component is non-zero for all collected samples, since otherwise there would be no component available to choose as a divisor \( x_i \). Such issues related to zeros in data is handled in 2.5.1.

The \( \text{alr} \) transform with its dimensional reduction inevitably introduces asymmetry between how components are treated. In some applications, this can negatively affect interpretation of results. The central log transform \( \text{clr} \) is designed to address this issue by transforming from \( \mathbb{S}^D \) to \( U_D \), where \( U_D = \{ (u_1, u_2, \ldots, u_d) : \sum_i u_i = 0 \} \). The \( \text{clr} \) transform thus maps the simplex to a hyperplane in \( \mathbb{R}^D \), retaining the same dimension as the original data.

**Definition 6.** Let \( \mathbf{x} = (x_1, x_2, \ldots, x_D) \in \mathbb{S}^D \). The *central log transform* \( \text{clr} : \mathbb{S}^D \to \mathbb{R}^D \) is defined by

\[
\text{clr}(\mathbf{x}) = \left( \log \left( \frac{x_1}{g(\mathbf{x})} \right), \log \left( \frac{x_2}{g(\mathbf{x})} \right), \ldots, \log \left( \frac{x_D}{g(\mathbf{x})} \right) \right)
\]

where \( g(\mathbf{x}) \) denotes the geometric mean of \( \mathbf{x} \).

Though the \( \text{clr} \) transform handles the problem of asymmetry introduced by the \( \text{alr} \) transform, immediate drawbacks are present. Most notably, any presence of zero values in the data is detrimental, since the geometric mean of such data will always be zero. A similar problem can occur when the number of data samples is very large. Suppose we have a \( k \) part composition, \( 1 < i < k \) and \( N \) samples of this composition, \( 1 < j < N \). Then, since each \( x_{ij} < 1 \), the product \( x_{i1} \times x_{i2} \times \ldots \times x_{iN} \) rapidly approaches zero for large \( N \). This can lead to numerical problems when using computational software such as R or Matlab. This can be addressed by choosing a smaller subset of data to be analysed. Another drawback of the \( \text{clr} \) transform is the fact that while it keeps the dimensionality of the compositional data intact it also retains the linear dependence in the components, leading to a singular covariance matrix. This method would then be of limited use in any statistical analysis where such properties are detrimental such as linear regression.

The final transform under consideration is the *isometric log transform*. The purpose of this transform is to deal properly with compositional covariates, and to also address the issue of singular covariance matrices which arises when using the \( \text{clr} \) transform, while retaining a sense of symmetry not present in the \( \text{alr} \) transform. This transform, and its implementation is described thoroughly in [9], and is described in terms of an orthonormal basis in \( \mathbb{R}^{D-1} \).

**Definition 7.** Let \( \mathbf{x} = (x_1, x_2, \ldots, x_D) \in \mathbb{S}^D \), and \( \mathbf{H} \) be a \( D \times (D - 1) \) matrix whose columns form an orthonormal basis in \( \mathbb{R}^{D-1} \). The *isometric log transform* \( \text{ilr} : \mathbb{S}^D \to \mathbb{R}^{D-1} \) is defined by

\[
\text{ilr}(\mathbf{x}) = \mathbf{H}^T \text{clr}(\mathbf{x})
\]

The inverse transform is considerably more complicated than previous cases and will be omitted, for more information see [9].
2.5.1 Handling data with zero values

Each of the proposed transforms has difficulties with zero data. The alr transform is the most robust in this sense, since if the user has a data set where at least one component is always non-zero, one can choose this component as the divisor $x_i$ in (4). Here we discuss some methods for handling such problems in general.

To handle the issue of zero values in a simple and easily implemented way, one can choose a $\delta > 0$ suitably small to replace each zero value, and then have each measurement vector be divided through with the component sum. The resulting vector remains in $S^D$ as a slightly distorted version of the original data, and from this point any of the above transforms can be successfully applied. This distortion of the data could however influence the predictive power of the data, which is something that the user should take into account. A version of this technique is defined using the additive perturbation operator, which we define now:

**Definition 8.** The additive perturbation operator which we denote $\circ$ is defined for a vector $x \in S^D$ and a vector $y \in \mathbb{R}^D$ by

$$z = x \circ y = C(x_1 + y_1, x_2 + y_2, ..., x_D + y_D)$$

with inverse w.r.t $x$ given by

$$x = z \circ^{-1} y = z \circ (-y) = (z_1 - y_1, z_2 - y_2, ..., z_D - y_D) \times (1 + \sum_{i=0}^D y_i)$$

From this we can formalise the above method of handling zero data. Let $\delta > 0$ be suitable chosen and $D = (\delta, \delta, ..., \delta)$ be a vector of length $D$. Suppose $x$ is a compositional measurement vector containing zeros. Then $x \circ D = y$ is again a compositional vector, but with all components non-zero. After analysis has been performed on $y$, the results can be restored to zero containing form by the inverse transform given in the definition.

2.6 Linear Regression In The Simplex

In developing the theory of multivariate regression for data in $S^D$, there are some details that need clarification. In performing linear regression on compositional data, the transformation methods of 2.5, in particular the alr transform, gives us the possibility to take the data into traditional real space, and estimation of the coefficient matrix $\beta$ and other parameters can be made in the standard environment of $\mathbb{R}^{D-1}$ of multivariable analysis. This is made rigorous in the next section. The fundamentally linear dependent nature of compositional covariates may render the estimate for the coefficient matrix poorly defined. In general, for covariates showing a strong correlation, the estimate of individual coefficients may be skewed. In these situations, one option is to exclude one variable from analysis since it does not provide any additional information. Otherwise, for data where strong multicollinearity is present it has been proposed [6, section 5.4.1, p 85] that large sample sizes can mitigate the inaccuracy in the estimation of the coefficient matrix. Another approach, unique to the simplical data case is explored in [7], whereby the author translates the standard theory of multivariate linear regression to the simplex environment. This approach bypasses the problem of linear dependence in the covariates by utilizing the geometry of the simplex, as opposed to the standard matrix methods of multivariate analysis. In this work we will be adopting the former approach.

2.6.1 Transformed Regression & The Logistic Normal Distribution

The nature of performing multivariate multiple linear regression on data transformed from the simplex into real space is completely analogous to standard multivariate analysis. For compositional data we can thus make the following definition.

**Definition 9.** Let $X$ be a a set of $N$ independent $D$-part compositions. Then $X$ is said to follow a logratio linear model $m$ if the transformed logratio data matrix $Y$ can be expressed in the form

$$Y = \Theta Z + E$$
where the covariate matrix $Z$, of order $N \times p_m$ and full rank $p_m$ is a matrix of known constants, the parameter matrix $\Theta$ is of order $p_m \times d$ and the $N \times d$ error matrix $E$ is assumed to consist of independent row vectors, each distributed as $\mathcal{N}^d(0, \Sigma)$.

Under this model, the methods from 2.1 become applicable in finding the MLE estimate for $\Theta$. 
2.7 Confidence Intervals, Resampling, & Bootstrap Methods

Considering point estimates, whether of a parameter of a distribution or the future value of a random variable for a given predictor, there is often a need to develop confidence and prediction intervals of obtained estimates. These interval statistics are determined by the distribution of the estimators used to estimate the parameter. In practice this is rarely known, therefore there is a rich theory for estimating the distributions of such parameters for use in inference and constructing interval statistics. The main problem of interest is then how to obtain an accurate estimate of the parameter distribution. One particularly useful way to estimate this sample distribution using only a set of observations is to use resampling methods such as the bootstrap, detailed in [10].

Bootstrap methods are based on the premise that a data set of size \( N \) has been obtained, and some statistic \( \Theta \) to be estimated, as a function of this data. Formally, the data \( x \) is a set of observations of a random variable \( X \), and we seek to find an estimate \( \hat{\Theta}(x) \) as an observation of an estimator, \( \Theta(X) \). The value of such an estimate can be calculated for several common estimators, such as the sample mean, sample variance, and the median. But without knowledge of the underlying distribution of \( X \), it is difficult to take the analysis further, for instance in constructing confidence intervals for the estimate \( \Theta \). Since \( \Theta \) is a function of a random variable \( X \), it is itself a random variable with some distribution, called \( \mathcal{F} \). We then have the property that the empirical distribution function for a sample of size \( n \), \( F_n^* \), can give us estimates of the properties of \( \Theta \), even though they cannot be observed. The process works as thus:

Suppose we have an estimate \( \hat{\Theta} \), with an unobservable true value of \( \Theta \). It may not always be the case that the distribution of \( \hat{\Theta} \) provides an accurate description of \( \Theta \). In constructing a confidence interval for \( \Theta \), we want to know how much the estimator \( \Theta \) varies with \( X \). We can then try to estimate the distribution of

\[
\Delta = \hat{\Theta} - \Theta.
\]

If the distribution of \( \Delta \) was known, we could find \( \delta_1, \delta_2 \) such that for a given value of \( \Theta \) we have

\[
P(\delta_1 \leq \Delta \leq \delta_2 \mid \Theta) = 1 - \alpha
\]

for any chosen level of significance \( \alpha \), which would provide us with a confidence interval for \( \Theta \) in the form

\[
[\hat{\Theta} - \delta_1, \hat{\Theta} + \delta_2].
\]

The question then becomes how to estimate the distribution of \( \Delta \). A key property [11] of the empirical distribution function \( \Delta^* \) is that

\[
\Delta^* \xrightarrow{\text{a.s.}} \Delta, \quad \text{as} \quad n \to \infty
\]

and by using the techniques of Monte Carlo simulations, this distribution \( \Delta^* \) can be approximated to arbitrary precision [11].

The idea is using resampling from our observational matrix \( x \). This process consists of using our data set \( x \) of size \( n \), and creating a new sample set \( x^* \), by using sampling with replacement from \( x \), choosing \( n \) such sample data points from \( x \). Then by calculating two estimates, \( \hat{\Theta} \) from the original data set \( x \), and a new estimate \( \hat{\Theta}^* \) from our resampled data set \( x^* \) we can estimate \( \Delta \) as \( \Delta^* = \hat{\Theta}^* - \hat{\Theta} \). Repeating this process, we obtain more and more point estimates for the distribution of \( \Delta^* \), and so an approximation of the distribution of \( \Delta \) is obtained that can be used for creating a confidence interval. After repeating the above process \( N \) times, we obtain \( N \) point estimates, which can then be ordered from smallest to largest

\[
\{\Delta^*_1 \leq \Delta^*_2 \leq \ldots \leq \Delta^*_N\}
\]

to obtain the empirical distribution of the parameter \( \Delta \). Computing \( i = \alpha \times N \) and \( j = (1 - \alpha) \times N \) (rounding if necessary) we get our \( \alpha \) level confidence interval as

\[
[\hat{\Theta} - \Delta^*_i, \hat{\Theta} + \Delta^*_j].
\]
The above method for developing confidence intervals of estimations is often referred to as the percentile bootstrap approach and is detailed more thoroughly in [8]. In this work, we seek to estimate \((1 - \alpha)\%\) confidence bands as well as prediction intervals for our regression curves using a heuristic single variable approach. Let \(X, Y\) be covariate matrix and response variable respectively, assumed to have a linear relationship in some coefficient matrix \(\beta\). Let \(X_p\) and \(Y_p\) be observations made. By sampling pairs \((X_i, Y_i)\) with replacement from \(X_p\) and \(Y_p\), an estimate for the coefficients \(\beta^*\) can be calculated by either standard multivariate regression as in 2.1 or through the transformation techniques discussed in 2.5, and then a prediction estimate can be made in the form \(Y^*_p = X_p\beta^*\). This provides us with a point estimate matrix of the components of \(Y_p\). Repeating this process \(N\) times, one obtains \(N\) different point estimates for each point along a regression curve. These can then be ordered and \(\alpha, 1 - \alpha\) percentiles can be calculated for each point estimate. Together, these point estimates provide a confidence band for the least squares fitted model of the data. To summarize, the algorithm for developing confidence bands for linear regression point estimates is

1. For \(R\) replicates, \(r = 1, 2, ..., R\), do
   
   (a) Draw with replacement sample pairs \((X_i, Y_i)\) from \(X, Y\).
   
   (b) From these pairs, estimate \(\beta^*\) and calculate \(Y^*_r = X_p\beta^*\).

2. For each point \((x_i, y_i)\), obtain \(R\) estimates. Sort these in ascending order and calculate \(\alpha\) and \(1 - \alpha\) percentiles of these replicates.

3. Obtain \(1 - \alpha\) confidence interval as

   \[ [(X^*_p\beta^*_\alpha, (X^*_p\beta^*)_{1-\alpha}) \]

   Confidence interval for a regression curve gives a bound on the behaviour of the estimated regression curve \(X\beta^*\). If we instead are interested in modelling the behaviour of \(Y^*\) itself around \(X\beta^*\), that is to develop prediction intervals for regression curves, an estimate of the noise induced variability of \(Y\) is needed. One approach is to resample among the residuals of the fitted model. This approach is handled thoroughly in [10].

The problem is to generate a prediction interval for a response variable \(Y_{N+1}\). Now, using the methods of 2.1 we can find an estimate \(\beta^*\) and use as a point predictor \(Y^*_{N+1} = X_{N+1}\beta^*\). The error made by such an estimate is

\[ \epsilon^*_{N+1} = Y_{N+1} - Y^*_{N+1} \]

which implies

\[ Y_{N+1} = \epsilon^*_{N+1} + Y^*_{N+1} \]

Since \(Y^*_{N+1}\) is known, one approach to bound \(Y_{N+1}\) would be to estimate the percentiles of \(\epsilon^*_{N+1}\) and form the interval

\[ [Y^*_{N+1} + \epsilon^*_{N+1,\alpha}, Y^*_{N+1} + \epsilon^*_{N+1,1-\alpha}] \]

To estimate the distribution of the errors, notice that

\[ \epsilon^*_{N+1} = Y_{N+1} - Y^*_{N+1} = X_{N+1}\beta + \epsilon_{N+1} - X_{N+1}\beta^* = X_{N+1}(\beta - \beta^*) + \epsilon_{N+1}. \]

By setting \(\beta = \beta^*\), that is using our least squares estimate of \(\beta\) as the ‘true’ value, and then, using the methods discussed above, generate new estimates \(\tilde{\beta}_p\), generate a new residual \(\epsilon_{N+1,p} = Y^* - X_{N+1} \tilde{\beta}_p\), we can obtain a point estimate for \(\epsilon^*_{N+1}\) as

\[ \epsilon^*_{N+1} = X_{N+1}(\beta^* - \tilde{\beta}_p) + \epsilon_{N+1,p} \]
Before we summarize the previous method, there is the practical issue to be addressed that the residuals often have small variances, and that a more stable estimate should be used. Therefore we will instead employ the variance corrected residuals, that is \( \{s_1 - \bar{s}, s_2 - \bar{s}, ..., s_N - \bar{s}\} \) where \( s_i = \frac{\tilde{\epsilon}_i}{\sqrt{h_i}} \) and \( h_i \) is the leverage of observation \( i \). With this out of the way, we summarize the above algorithm for generating bootstrap prediction intervals.

1. Use the methods of 2.1 to predict \( Y_{N+1}^* = X_{N+1} \beta^* \).
2. Calculate residuals \( E = \{\epsilon_1, \epsilon_2, ..., \epsilon_N\} = Y_N - X_N \beta^* \) and calculate the variance adjusted residuals \( S \).
3. For \( P \) resamples, \( p = 1, 2, .., P \), do
   (a) Draw a size \( N \) resample from \( S \) in step 2.
   (b) Generate \( Y^* = X_{N+1} \beta^* + \epsilon^* \).
   (c) Regress from \((X_{N+1}, Y^*)\) to obtain an estimate \( \tilde{\beta}_p \).
   (d) Obtain bootstrapped residual from this replication \( \epsilon_p = Y_N - Y^* \).
   (e) Calculate the variance adjusted residuals from \( \epsilon_p \).
   (f) Choose randomly an element of \( \epsilon_p \), call this \( \epsilon_p^* \).
   (g) Calculate this replications draw as \( \tilde{\epsilon}_p = X_{N+1} (\beta^* - \tilde{\beta}_p) \).
4. Find the \( \alpha \) and \( 1 - \alpha \) percentiles of these replicates.
5. Obtain prediction interval as

\[
\left[ Y_{N+1}^* + \epsilon_{\alpha}^*, Y_{N+1}^* + \epsilon_{1-\alpha}^* \right]
\]
3 Methods

3.1 Approach To Applying Linear Regression In The Simplex

In this work linear simplical regression was approached through the transformation techniques of 2.5, as well as the application of standard multivariate regression techniques discussed in 2.1 with results evaluated from each of the applied methods. For the transformation approach, the following paradigm for data analysis was adopted:

The Transformation Approach
For compositional dependent vectors $Y$ and (possibly compositional) covariates $X$

1. Inspect $Y$ for zero components, if such is found apply the techniques of 2.5.1.
2. Apply a chosen transform $f$ from 2.5 to $Y$.
3. Using a linear model of the form $f(Y) = Z = \beta X + E$ perform standard multivariate regression to find an estimate $\beta^*$ for this model.
4. Calculate the fitted model as $Z^* = \beta^* X$
5. Apply the inverse $f^{-1}$ to $Z^*$ to obtain a compositional fit $Y^*$.

After this procedure has been executed, there are several options for analysis. One such approach, adopted throughout most of this work, is the univariate approach in investigating the behaviour of individual components of a compositional distribution. For such components, familiar concepts from multivariate linear regression models such as $SSE$, $SST$, $R^2$ and residual analysis can all be employed to examine the goodness of fit for the estimated model. Recalling the familiar definition of the $R^2$ statistic for a random variable $Y$

$$R^2_j = 1 - \frac{SSE_j}{SST_j}$$

where $SST_j = \sum_{i=0}^{N}(Y_{ij} - \bar{Y}_j)^2$, with $\bar{Y}_j$ denoting the sample mean of $Y_j$ and $SSE_j = \sum_{i=0}^{N}(Y_{ij} - Y_{ij}^*)^2$, the sum of squared residuals. This statistic can then be calculated for each component of a compositional fit $Y^*$ to provide an insight into the individual fit of the component estimates. To judge the overall fit of the model, the following statistic is proposed:

$$R^2 = 1 - \sum_{i=0}^{D} \frac{SSE_i}{SST_i}$$

This statistic, which we will call total $R^2$ behaves similarly to the standard $R^2$, in that if each component $j$ has an individual $R^2_j$ close to 1, the total $R^2$ will be close to 1. Similarly, if each component has as low or negative $R^2_j$ value, the total will be low or negative. As opposed to something like the average component $R^2$, the total $R^2$ punishes deviating components more. If one component deviates in quality of fit it will reduce the total $R^2$ more than the average would, in that if some $R^2_j$ value decreases by some positive quantity $\alpha$, then the average decreases by $\frac{\alpha}{D}$ while the total $R^2$ decreases by $\alpha$.

An alternative definition, utilizing the simplex geometry, can be given by replacing the sample mean of $Y$ with the corresponding component of the centre of $Y$ as defined in 2.4, that is letting $\text{Cen}(Y) = (\hat{y}_1, \hat{y}_2, \ldots, \hat{y}_D)$

$$R^2_j = 1 - \frac{\sum_{i=0}^{N}(y_{i} - y_{ij}^*)}{\sum_{i=0}^{N}(y_{i} - \hat{y}_j)}$$

and similarly these values can be summed across all components to obtain an overall measure of fit of the model.
\[ R^2 = 1 - \sum_{j=1}^{D} \frac{\sum_{i=0}^{N} (y_i - y_i^*)}{\sum_{i=0}^{N} (y_i - \bar{y})} \]

Both of these measures were utilized in order to judge the fit of models.

### 3.2 Simulations Of Compositional Data

The behaviour of regression analysis on compositional data was analysed by constructing a simple model of such data and performing statistical analysis. The main point of interest was the difference in the predictive strength of applying standard multivariate analysis to the untransformed components in contrast to the transformed data under a suitable transform, described in 2.5.

The data was simulated using the following paradigm: Simulate \( k \) variables \( X_i, 1 \leq i \leq k \) sampled from a chosen probability distribution, for a composition of \( D \) parts construct \( N \) dependent variables \( Y_j \) as follows

\[
Y_1 = a_{11} + b_{12}X_1 + \cdots + b_{1k}X_k + \epsilon_1 \\
Y_2 = a_{21} + b_{22}X_1 + \cdots + b_{2k}X_k + \epsilon_2 \\
\vdots \\
Y_n = a_{n1} + b_{n2}X_1 + \cdots + b_{nk}X_k + \epsilon_n
\]

where the errors \( \epsilon_i \in E \sim N(0, \Sigma) \). From this simulate the observation of a compositional random variable by computing

\[
\tilde{Y} = C (Y_1, Y_2, \ldots, Y_D)
\]

and consider \( \tilde{Y} \) as a (non linear) function of the predictors \( X_i \) or in matrix notation

\[
\tilde{Y} = \Theta X + \tilde{E}
\]

where the first column vector of \( X \) is all ones. Repeating this process \( n \) times, an \( n \) sized sample is obtained of a compositional variable \( \tilde{Y} \). In the simulation of components, it is notable that each simulated \( Y_i \) are each required to be non-negative. This places some restrictions on the choice of covariance matrix \( \Sigma \) for the simulated errors. This can be addressed by choosing variances \( \sigma_i \) such that \( 3\sigma_i < \sum_{i=0}^{D} b_i a_i \), which then constrict possible negative outputs to less than 1% of the simulated vector components.

#### 3.2.1 Types Of Simulated Distributions

One of the main objects of interest in this study are skew distributions, where each observation has components that is significantly different in magnitude than the others. An example of this could be an observation of a 3 part composition where one component is 2%, another one is at 3% and the last one is 95%. The robustness of each method of analysis can then be tested against variations of such data to determine if the extremes are handled significantly different in either method. A variation of the skew compositional distributions are distributions containing one or more zero values. Since none of the transformation methods discussed in 2.5 easily handle such data we apply the methods discussed in section 2.5.1.

#### 3.2.2 Simulated Distributions

**Model 1**

A multivariate compositional distribution with three independent variables \( X_1, X_2, X_3 \), each following a normal distribution was considered; each \( X_i \sim \mathcal{N}(10, 1) \) and the data generated under the paradigm described above. The coefficient matrix \( \beta \) for the covariates and the error covariance matrix \( \Sigma \) for the simulated errors were
\[
\beta = \begin{bmatrix}
3 & 14 & 7 \\
2 & 6 & 9 \\
0.5 & 3.2 & 5
\end{bmatrix},\quad \Sigma = \begin{bmatrix}
10 & 5 & 5 \\
5 & 10 & 5 \\
5 & 5 & 10
\end{bmatrix}
\]

First a single distribution was generated with sample size \( n = 50 \), and goodness of fit statistics where calculated. Then using the methods of 2.1 and 3.1 two models where fitted, hereby referred to as the \textit{mvr} fit and \textit{alr} fit respectively. Subsequently, \( N = 1000 \) such distributions were generated and the resulting goodness of fit statistics averaged over all the simulations for the fitted models.

\textit{Model 2}

The second simulated data set was set up as using an uniformly distributed covariate \( X \sim U(1, 2) \); the \( \beta \) and \( \Sigma \) matrices used was

\[
\beta = \begin{bmatrix}
12 & 100 & 13 \\
12 & 0.01 & 32
\end{bmatrix},\quad \Sigma = \begin{bmatrix}
5 & 2.5 & 2.5 \\
2.5 & 5 & 2.5 \\
2.5 & 2.5 & 5
\end{bmatrix}
\]

For both the \textit{mvr} and \textit{alr} estimated models, a logarithmic fit was used. As in the first case, first a single realisation of size \( n = 50 \) was generated and analysed, and then \( N = 1000 \) such distributions were generated with goodness-of-fit statistics averaged across each realisation. Two models were fitted, one with the \textit{mvr} approach and one with the \textit{alr}.
3.3 Case Study: Industrial Processing

In this case study we are considering the outcome of an industrial process. The process produces a distribution of rocks of various sizes, classified by their diameter, which is then analysed by use of laser topography scanning, along with analytical algorithms to estimate the distribution. In addition, during the process the rocks are transported along a conveyor belt and through a drop, where an accelerometer records the impact of the falling rocks before they are transported further. The problem is then to estimate the original relative size class distribution of the rocks, using the overhead laser scanning and associated estimation algorithms and the accelerometer impulse data using both standard multivariate analysis and the application of the \( \text{alr} \) transform regression techniques. Figure 1 illustrates the situation.

Figure 1: An industrial process where rocks are transported along a conveyor belt, put through a reloading stage, and then scanned by a laser.

The laser scanner and associated algorithm estimates the size distribution of a rock sample, both in terms of area and volume. A size range has been set, classifying rocks on the basis of their cross-sectional diameter. The size ranges of the rock samples have been set (in mm), eight in total, as shown below.

\[
[0.0040, 0.0070, 0.0150, 0.0220, 0.0300, 0.0370, 0.0440, 0.0740]
\]

To illustrate the nature of the data, an excerpt of the response data is displayed in table 1.
The first size class is zero throughout all the samples, so this component may be omitted from analysis. Other components are occasionally zero as well, this indicates that the data must be modified slightly using the techniques discussed in section 2.7 in order for transformation techniques to be applicable. The covariates to be analysed are

- $X_1$: Laser Topography Area Data
- $X_2$: Mean/Max Impulses
- $X_3$: Mean Impulses
- $X_4$: Laser Topography Volume Data

Here the information provided by the area and volume estimates are believed to provide a better insight on the nature of the distribution of the rocks, while the impulses are considered to provide auxiliary information to boost the accuracy offered by the fitted models. In validating our model, we will set aside a training set and a verification set of the data. The goodness-of-fit statistics will be calculated from the verification set.

In analysing the industrial data of rock distributions, the covariates are set up and tested according to the following design scheme:

Table 2: A sample of 20 data points from the measured distribution of rock samples

<table>
<thead>
<tr>
<th>Rock Distribution Data</th>
</tr>
</thead>
<tbody>
<tr>
<td>(0, 0.6071, 0.1501, 0.0890, 0.0461, 0.0740, 0.0337, 0)</td>
</tr>
<tr>
<td>(0, 0.6434, 0.1617, 0.0795, 0.0999, 0.0245, 0, 0)</td>
</tr>
<tr>
<td>(0, 0.6236, 0.1592, 0.1072, 0.0667, 0.0432, 0, 0)</td>
</tr>
<tr>
<td>(0, 0.5335, 0.2076, 0.1043, 0.0863, 0.0682, 0, 0)</td>
</tr>
<tr>
<td>(0, 0.5106, 0.1900, 0.0864, 0.0873, 0.0403, 0.0357, 0.0497)</td>
</tr>
<tr>
<td>(0, 0.5136, 0.1958, 0.0707, 0.0516, 0.0232, 0.1029, 0.0422)</td>
</tr>
<tr>
<td>(0, 0.5257, 0.1921, 0.0733, 0.0855, 0.0238, 0, 0.0995)</td>
</tr>
<tr>
<td>(0, 0.5552, 0.2039, 0.0877, 0.0696, 0.0290, 0, 0.0547)</td>
</tr>
<tr>
<td>(0, 0.4792, 0.2304, 0.0962, 0.0816, 0.0661, 0.0465, 0)</td>
</tr>
<tr>
<td>(0, 0.4595, 0.2102, 0.0744, 0.0958, 0.0693, 0.0168, 0.0740)</td>
</tr>
<tr>
<td>(0, 0.5207, 0.2299, 0.1057, 0.0635, 0.0395, 0.0407, 0)</td>
</tr>
<tr>
<td>(0, 0.3911, 0.1848, 0.0530, 0.0991, 0.0131, 0.0804, 0.1785)</td>
</tr>
<tr>
<td>(0, 0.4819, 0.2195, 0.0752, 0.0729, 0.0753, 0, 0.0753)</td>
</tr>
<tr>
<td>(0, 0.6111, 0.1785, 0.0898, 0.0816, 0.0390, 0, 0)</td>
</tr>
<tr>
<td>(0, 0.5895, 0.1765, 0.0886, 0.0587, 0.0587, 0.0281, 0)</td>
</tr>
<tr>
<td>(0, 0.6061, 0.1849, 0.0778, 0.0878, 0.0434, 0, 0)</td>
</tr>
<tr>
<td>(0, 0.5765, 0.1875, 0.0855, 0.0949, 0.0283, 0.0273, 0)</td>
</tr>
</tbody>
</table>

Table 3: Array of designs used in the case study.

<table>
<thead>
<tr>
<th>Design Matrix</th>
<th>$X_1$</th>
<th>$X_2$</th>
<th>$X_3$</th>
<th>$X_4$</th>
</tr>
</thead>
<tbody>
<tr>
<td>D1</td>
<td>×</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>D2</td>
<td>×</td>
<td>×</td>
<td></td>
<td></td>
</tr>
<tr>
<td>D3</td>
<td>×</td>
<td>×</td>
<td>×</td>
<td></td>
</tr>
<tr>
<td>D4</td>
<td></td>
<td>×</td>
<td></td>
<td></td>
</tr>
<tr>
<td>D5</td>
<td></td>
<td>×</td>
<td>×</td>
<td></td>
</tr>
<tr>
<td>D6</td>
<td></td>
<td>×</td>
<td>×</td>
<td>×</td>
</tr>
<tr>
<td>D7</td>
<td></td>
<td>×</td>
<td>×</td>
<td>×</td>
</tr>
<tr>
<td>D8</td>
<td></td>
<td>×</td>
<td></td>
<td>×</td>
</tr>
<tr>
<td>D9</td>
<td></td>
<td>×</td>
<td>×</td>
<td>×</td>
</tr>
<tr>
<td>D10</td>
<td></td>
<td>×</td>
<td>×</td>
<td>×</td>
</tr>
</tbody>
</table>
In the definition of the total covariance matrix $\Gamma$, it is clear that any zero components in the data will result in undefined values. There are two apparent solutions to this problem. First, one could apply the zero correction scheme from section 2.5.1 to the response data, which would then make calculation of the total covariance matrix possible. A second approach would be to omit all measurement vectors containing zero data, in the hope that there will be sufficiently many left for satisfactory analysis. The second option is clearly favourable, if the size of the non-zero measurements is large enough to be representative of the overall population. In this case the dependent matrix has a total of 280 measurements, out of which 79 contained no zeros (other than the constant zero component 1), which is sufficient for applying the second alternative to these measurements directly.
4 Results

4.1 Simulation Study

4.1.1 Simulation 1: 3-Component Multivariate Normal Distribution

On a single data set from simulation 1, both the \textit{mvr} and the \textit{alr} provided a close fit to the simulated data. Both models had close fitting curves, and high $R^2$ values of all types to match, see figure 2 and table 4 respectively.

![Figure 2: A simulation of \(n = 50\) data points of compositional data generated using the methods described in 3.2.1 with three independent covariates all following a normal distribution with normally distributed errors. Blue lines indicate simulated data points, the red lines indicate the mvr fitted model, and the green denotes the alr model. Magenta denotes the 95\% confidence band for the mvr estimate, while black denotes the corresponding estimate for the alr fit.](image)

Notably, there seems to be little difference in the closeness of the 95\% CI, with the \textit{mvr} CI being slightly wider; notably before \(x = 5\) in each simulation.

<table>
<thead>
<tr>
<th>Regression Stats:</th>
<th>mvr fit</th>
<th>alr fit</th>
</tr>
</thead>
<tbody>
<tr>
<td>SS</td>
<td>0.0009887</td>
<td>0.0009082</td>
</tr>
<tr>
<td>Component $R^2$</td>
<td>(0.9695 0.8957 0.9709)</td>
<td>(0.9727 0.8944 0.9732)</td>
</tr>
<tr>
<td>Total $R^2$</td>
<td>0.8361</td>
<td>0.8402</td>
</tr>
<tr>
<td>Component $\mathcal{R}^2$</td>
<td>(0.9695 0.8957 0.9710)</td>
<td>(0.9727 0.8944 0.9732)</td>
</tr>
<tr>
<td>Total $\mathcal{R}^2$</td>
<td>0.8362</td>
<td>0.8403</td>
</tr>
</tbody>
</table>

When investigating the residuals, there is no clear difference between the two methods as can be seen...
in figure 3. The residual variance is both small in value in relation to the data, and there is no significant difference between the two methods, see table 5.

Figure 3: Component-wise plot of the residuals of simulation 1, red indicating the residuals form the mvr fit and the green is the residuals from the alr fit.

Table 5: Sample variance of the residuals from a single realisation of simulation 1

<table>
<thead>
<tr>
<th>Variance</th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>mvr fit</td>
<td>1.0e-04 × (0.1071, 0.0419, 0.0527)</td>
</tr>
<tr>
<td>alr fit</td>
<td>1.0e-04 × (0.09749, 0.04199, 0.04587)</td>
</tr>
</tbody>
</table>
Multiple repetitions of simulation 1 were performed ($N = 1000$); the above observations were shown to be consistent across many simulations, with the averaged total sum of square errors and $R^2$ values given in Table 6 being consistently low and high, respectively. The results are then found to be consistent for the model in simulation 1.

<table>
<thead>
<tr>
<th>Regression Stats:</th>
<th>mvr fit</th>
<th>alr fit</th>
</tr>
</thead>
<tbody>
<tr>
<td>SS</td>
<td>0.0011</td>
<td>0.001</td>
</tr>
<tr>
<td>Component $R^2$</td>
<td>(0.9422 0.8866 0.9756)</td>
<td>(0.9568 0.9010 0.9399)</td>
</tr>
<tr>
<td>Total $R^2$</td>
<td>0.8124</td>
<td>0.8701</td>
</tr>
<tr>
<td>Component $R^2$</td>
<td>(0.9422 0.8866 0.9756)</td>
<td>(0.9568 0.9010 0.9399)</td>
</tr>
<tr>
<td>Total $R^2$</td>
<td>0.8124</td>
<td>0.8702</td>
</tr>
</tbody>
</table>

4.1.2 Simulation 2: 3-Component Univariate Uniform Distribution

In simulation 2 the alr method provides a much better fit to the data than the mvr method as seen in figure 4. In particular, the mvr underestimates the early steep curvature of the data, and subsequently fails to predict the stabilization of the curves after $x = 25$.

Figure 4: A simulation of $n = 50$ data points of compositional data generated using the methods described in 3.2.1 with a single independent variable all following a uniform distribution with normally distributed errors. Blue lines indicate simulated data points, the red lines indicate the mvr fitted model, and the green denotes the alr model.

In figure 5, the same regression curves are on display with the added confidence bands of the point estimates which show again the alr method having a narrower band; indicating that the model provided by the alr is more consistent in the point estimates of the regression curves; though neither is too far off.
Figure 5: A simulation of $n = 50$ data points of compositional data generated using the methods described in 3.2.1 with a single independent dependent variable all following a uniform distribution with normally distributed errors. Blue lines indicate simulated data points, the red lines indicate the mvR fitted model, and the green denotes the alr model. Magenta denotes the 95% confidence band for the mvR estimate, while black denotes the corresponding estimate for the alr fit.

The regression statistics are displayed in table 5, with component $R^2$ values being high all around, which would indicate a very good fit for both models; yet the overall $R^2$ values are by a significant margin in favour of the alr estimate. This corroborates what was previously visible, that the alr model provides a better fit.

Table 7: Regression statistics from a single realisation of simulation 2

<table>
<thead>
<tr>
<th>Regression Stats</th>
<th>mvR fit</th>
<th>alr fit</th>
</tr>
</thead>
<tbody>
<tr>
<td>SS</td>
<td>0.0009887</td>
<td>0.0009082</td>
</tr>
<tr>
<td>Component $R^2$</td>
<td>(0.9143 0.9421 0.9477)</td>
<td>(0.9886 0.9984 0.9982)</td>
</tr>
<tr>
<td>Total $R^2$</td>
<td>0.8041</td>
<td>0.9852</td>
</tr>
<tr>
<td>Component $R^2$</td>
<td>(0.9220 0.9452 0.9500)</td>
<td>(0.9896 0.9984 0.9983)</td>
</tr>
<tr>
<td>Total $R^2$</td>
<td>0.8173</td>
<td>0.9864</td>
</tr>
</tbody>
</table>

The residuals in figure 6 clearly show non-random tendencies, in particular for the mvR estimate. The alr estimate shows slight non-stochastic behaviour before $x = 5$ then stabilizes around 0 as was expected. Further the residuals of the mvR fit has an order of magnitude greater dispersion than the alr; see table 9.
Figure 6: Component-wise plot of the residuals of simulation 2, red indicating the residuals form the mvr fit and the green is the residuals from the alr fit.

Table 8: Sample variance of the residuals from a single realisation of simulation 2

<table>
<thead>
<tr>
<th>Variance Of Residuals:</th>
<th>mvr fit</th>
<th>alr fit</th>
</tr>
</thead>
<tbody>
<tr>
<td>mvr fit</td>
<td>1.0e-03 × (0.0443, 0.6636, 0.3873)</td>
<td></td>
</tr>
<tr>
<td>alr fit</td>
<td>1.0e-03 × (0.0138, 0.0120, 0.0112)</td>
<td></td>
</tr>
</tbody>
</table>

On repeated realisations of simulation 2 for \( N = 1000 \) times, the average regression statistics corroborates the results of the single realisation, showing this result to be consistent. The mvr estimate shows considerably higher sum of square errors, and the alr has higher overall \( R^2 \) values on all front, confirming that the alr provides a closer fit.

Table 9: Regression statistics averaged from 1000 realisations of simulation 2

<table>
<thead>
<tr>
<th>Regression Stats:</th>
<th>mvr fit</th>
<th>alr fit</th>
</tr>
</thead>
<tbody>
<tr>
<td>SS</td>
<td>0.0475</td>
<td>0.0029</td>
</tr>
<tr>
<td>Component ( R^2 )</td>
<td>(0.9081, 0.9160, 0.9001)</td>
<td>(0.9988, 0.9885, 0.9993)</td>
</tr>
<tr>
<td>Total ( R^2 )</td>
<td>0.7984</td>
<td>0.9797</td>
</tr>
<tr>
<td>Component ( R^2 )</td>
<td>(0.9982 0.9675 0.9800)</td>
<td>(0.9989, 0.9986, 0.9994)</td>
</tr>
<tr>
<td>Total ( R^2 )</td>
<td>0.8184</td>
<td>0.9968</td>
</tr>
</tbody>
</table>
4.2 Results of Case Study

The estimate of the $\Gamma$ matrix of the dependent vector is displayed below, with the upper triangle denoting the relative standard deviation of components and the lower triangle the expected log ratios of components. The blank entries corresponds to the zero component row in the data, and is included for easier reference indexing. The estimated $\Gamma^*$ matrix confirms the heuristic notion that component $Y_2$ is the largest on average, with $\xi_{i,1}$ positive for all $i$ and greater than $\tau_{i,1}$ for all $i$. The overall descending nature of the components in a composition is reflected in the positive expected values across the matrix. The negative entry in $\xi_{6,7}$, though close to zero, reflects how in the estimate component 6 appears smaller than 7 on average, although only slightly. The overall correlation between components is quite low, indicating no strong relationship between the behaviour of the components.

\[
\Gamma^* = \\
\begin{bmatrix}
0 & - & - & - & - & - & - \\
- & 0 & 0.3916 & 0.4548 & 0.4685 & 0.5840 & 0.7196 & 0.6301 \\
- & 0.5884 & 0 & 0.1489 & 0.2198 & 0.3957 & 0.6069 & 0.5636 \\
- & 1.4122 & 0.8238 & 0 & 0.2824 & 0.3749 & 0.6617 & 0.6220 \\
- & 1.5690 & 0.9806 & 0.1567 & 0 & 0.4748 & 0.6429 & 0.6093 \\
- & 2.0779 & 1.4895 & 0.6657 & 0.5089 & 0 & 0.8055 & 0.7298 \\
- & 1.9931 & 1.4048 & 0.5809 & 0.4242 & -0.0848 & 0 & 0.7720 \\
- & 2.1205 & 1.5321 & 0.7083 & 0.5515 & 0.0426 & 0.1273 & 0
\end{bmatrix}
\]

As visible in table 10, the design matrix with the highest $R^2$ values was $D1$, and the corresponding regression curves for the mvr and the alr fitted prediction models are displayed in figure 7 and 8.

<table>
<thead>
<tr>
<th>Design Matrix:</th>
<th>mvr $R^2$</th>
<th>mvr $R^2$</th>
<th>alr $R^2$</th>
<th>alr $R^2$</th>
</tr>
</thead>
<tbody>
<tr>
<td>D1:</td>
<td>0.7936</td>
<td>0.7959</td>
<td>0.7120</td>
<td>0.7199</td>
</tr>
<tr>
<td>D2:</td>
<td>0.6067</td>
<td>0.7827</td>
<td>0.3695</td>
<td>0.6851</td>
</tr>
<tr>
<td>D3:</td>
<td>0.5927</td>
<td>0.7737</td>
<td>0.3674</td>
<td>0.6827</td>
</tr>
<tr>
<td>D4:</td>
<td>0.6095</td>
<td>0.7864</td>
<td>0.3533</td>
<td>0.6770</td>
</tr>
<tr>
<td>D5:</td>
<td>0.5724</td>
<td>0.7532</td>
<td>0.4213</td>
<td>0.6909</td>
</tr>
<tr>
<td>D6:</td>
<td>0.5936</td>
<td>0.7702</td>
<td>0.4281</td>
<td>0.7013</td>
</tr>
<tr>
<td>D7:</td>
<td>0.5913</td>
<td>0.7694</td>
<td>0.4448</td>
<td>0.7078</td>
</tr>
<tr>
<td>D8:</td>
<td>0.6111</td>
<td>0.7844</td>
<td>0.4392</td>
<td>0.7084</td>
</tr>
<tr>
<td>D9:</td>
<td>0.6039</td>
<td>0.7804</td>
<td>0.4592</td>
<td>0.7154</td>
</tr>
</tbody>
</table>

In both cases the fit seems to highly dependent on which component is being estimated, as both models have similar performances for components $Y_2$ and $Y_3$ which are relatively large, but the predictions are less accurate for the components of smaller relative size. Here the noise of the data seems to be larger in comparison to the magnitude of the measurements, leading the fitted models to underestimate fluctuations in the data.
Figure 7: Component-wise fitted model using the mvr methods for the design matrix D1. Blue indicates measured data, and red indicates the mvr fitted model, with prediction bands given in green.
Neither method shows any real advantage in this regard. The regression statistics seem to back up this analysis, since the first four components which are fairly large in relative size are estimated well, but the last three which have values in the single digit range (many of them zero) are estimated poorly. In both these cases the centred statistic, $R^2$ shows higher values, both in components and total. Here the $mvr$ method comes out victorious, with larger values in both categories. A notable feature of the $alr$ method in components $Y_7$ and $Y_8$ is that it avoids predicting negative values, which the $mvr$ does.
Table 11: Regression statistics calculated from the mvr fitted model

<table>
<thead>
<tr>
<th>Regression Stats: mvr fit</th>
</tr>
</thead>
<tbody>
<tr>
<td>SSE</td>
</tr>
<tr>
<td>Component $R^2$</td>
</tr>
<tr>
<td>Total $R^2$</td>
</tr>
<tr>
<td>Component $R^2$</td>
</tr>
<tr>
<td>Total $R^2$</td>
</tr>
</tbody>
</table>

Table 12: Regression statistics calculated from the alr fitted model

<table>
<thead>
<tr>
<th>Regression Stats: alr fit</th>
</tr>
</thead>
<tbody>
<tr>
<td>SSE</td>
</tr>
<tr>
<td>Component $R^2$</td>
</tr>
<tr>
<td>Total $R^2$</td>
</tr>
<tr>
<td>Component $R^2$</td>
</tr>
<tr>
<td>Total $R^2$</td>
</tr>
</tbody>
</table>

The residuals for each fit, visible in figure 9, show that the alr fitted model has less dispersion yet seems less stochastic in nature, indicating that while the alr approach provides less overall error in this context the error is more systematic.

Figure 9: Component-wise residuals for the mvr fitted model (red), with the residuals for the alr fitted model (green)

Table 13: Component-wise sample variance of the residuals for the mvr and alr fitted models

<table>
<thead>
<tr>
<th>Variance :</th>
</tr>
</thead>
<tbody>
<tr>
<td>mvr fit</td>
</tr>
<tr>
<td>alr fit</td>
</tr>
</tbody>
</table>
5 Discussion & Conclusions

5.1 Simulation Study

Based on the results for simulation 1 there seems to be little difference in the two regression methods. This seems to result from the relatively low error variance in the simulated model, along with the similar and roughly constant magnitude of the components; so it seems that both types of regression models behave well for this type of data. One difference comes in the confidence bands, which are narrower for the alr model, especially in simulation 2, which indicates lower variation in the quality of point estimates; that the mvr with greater frequency provides more extreme estimates, while the alr is more stable. The reasons behind this mechanism would be of interest in future research, to gather more understanding of how regression models behave in the simplex context. In this simulation, there is notably no meaningful difference between the component and total $R^2$, either when using the centre based statistic $R^2$ or using the standard definition in of $R^2$; as in their values are very close. In comparing with simulation 2, where there were slight differences between the two measures and to the industrial data application where they differed significantly, it seems that the statistics differ when the data is less stable, i.e with greater noise, with the centred $R^2$ reporting higher values in this case.

In light of simulation 2, it seems that the main weakness of applying the mvr approach to compositional data may be found in non-linear functions between the dependent and independent variables.

Conclusions Of Simulations From the simulated distributions it seems like the standard multivariate analysis viewpoint can be satisfactory in preliminary analysis of the overall distribution. If the user is mainly interested the overarching behaviour of the distributions then standard multivariate analysis and a cumulative sum diagram can be surprisingly satisfactory. The standard methods would seem to be applicable if the data is relatively constant, without too much variation in response to the independent variables.

In analysing data with stronger trends, the mvr estimate seems unsuitable for making reliable predictions regarding the long-term behaviour of the model. Despite this, the $R^2$ statistics calculated (in particular with regards to simulation 2) would indicate that the mvr fitted model is quite satisfactory, when both the estimations of negative percentages and considerably weak estimation of the trend are both visible in figure 4.

Since the standard multivariate analysis methods have no restriction in regards to the simplex structure, standard regression techniques may estimate negative components, as well as components greater than unity. The erroneous analysis of standard regression techniques such as estimating negative components, as well as components greater than unity is not visible from cumulative sum diagrams, or necessarily for standard regression curves, so the practitioner should take great care in what measures to use when judging the quality of the fitted model. The simplicity and computational inexpensiveness of the alr transform approach, coupled with both strong theoretical motivation and improved results in several cases in terms of fit, indicates its importance and viability as an important tool of compositional analysis.

5.2 Industrial Data

It would seem that both methods perform similarly on the rock distribution data, with the alr avoiding negative estimates of small components at perhaps the expense of a closer fit. From the data it would seem that components $Y_7$ and $Y_8$ have a larger relative variance, which both methods have difficulties handling. It should be noted that due to the large number of zeros in the data, the alr transform has to be performed on adjusted data, which could affect the predictive power of the model. The alr model performs slightly worse in terms of $R^2$ values; this could then be explained by the former necessity to compensate for zero data, and in general the methods then perform similarly. It is worth noting that the standard $R^2$ statistics for the alr fit seem almost overly pessimistic in this context, with values lower than a visual inspection of the fits would imply. Now the centred $R^2$ is higher for these estimates, and seem to more accurately portray the quality of fit visible in figure 8. In either case there is clearly a difference in how the statistics judge the alr fitted model.
Notably the residuals for the \textit{alr} indicate some sort of systematic error with its non-random nature. Yet the smaller absolute variation could make the \textit{alr} a favourable choice for analysis under certain circumstances.

**Conclusions Of Case Study**  Similarly to the performance in the simulated data, standard multivariate analysis techniques, while mostly satisfactory in a broad stroke analysis, is inadequate under scrutiny. While the overall fit of the curve can be seen as satisfactory, as indicated by figure 7, the methods misrepresents the data under certain circumstances as in negative percentages. The \textit{alr} transform by design avoids these difficulties, and in general does this with similar accuracy of fit, see table 9.
6 Acknowledgements

This project was performed under the supervision of Patrik Ryden, who provided guidance and invaluable advice throughout the process. Thanks to Daniel Lindmark for providing the code for data handling of the industrial data, which made the practicalities of analysis orders of magnitude simpler. Special thanks to Andrè Berghlund and Erik Jörgenfält for providing insight and advice in times of uncertainty, and to Per Åhag who provided great moral support.

References


7 Appendix

7.0.1 Proof of Proposition 1, section 2.1

Proof. The proof is standard multivariate analysis. Consider the linear model \( m, (2) \), and the univariate component associated with the \( i \)th response variable with \( N \) measurements

\[
y_i = X\Theta_i + \epsilon_i
\]

\[
\begin{bmatrix}
y_1 \\
y_2 \\
\vdots \\
y_N
\end{bmatrix} =
\begin{bmatrix}
1 & x_{11} & \cdots & x_{k1} \\
1 & x_{12} & \cdots & x_{k2} \\
\vdots & \vdots & \ddots & \vdots \\
1 & x_{1N} & \cdots & x_{kN}
\end{bmatrix}
\begin{bmatrix}
\theta_1 \\
\theta_2 \\
\vdots \\
\theta_k
\end{bmatrix}
+ 
\begin{bmatrix}
\epsilon_1 \\
\epsilon_2 \\
\vdots \\
\epsilon_N
\end{bmatrix}.
\]

We seek to minimize the sum of squares of the residuals, that is \( \epsilon'\epsilon \) (dropping the sample index subscript \( i \) for notational simplicity):

\[
\epsilon' \epsilon =
\begin{bmatrix}
\epsilon_1 \\
\epsilon_2 \\
\vdots \\
\epsilon_N
\end{bmatrix}
\begin{bmatrix}
\epsilon_1 \\
\epsilon_2 \\
\vdots \\
\epsilon_N
\end{bmatrix} = \sum_{i=1}^{N} \epsilon_i^2
\]

By definition, we have \( \epsilon = y - X\Theta \) so that \( \epsilon'\epsilon = (y - X\Theta)'(y - X\Theta) \). After expansion, we are faced with the following minimization problem:

Minimize \( \Theta : \epsilon'\epsilon = y'y - 2\Theta'X'y + \Theta'X'X\Theta \) \hspace{1cm} (5)

Taking the derivatives with respect to \( \Theta \) and setting the result equal to zero, together with the following identities

\[
\frac{\partial (\Theta'X'y')}{\partial \Theta} = X'y
\]

\[
\frac{\partial (\Theta'X'X\Theta)}{\partial \Theta} = 2X'X\Theta
\]

we obtain the equation

\( X'X\Theta = X'y \)

with the obvious solution

\( \Theta = (X'X)^{-1}X'y \)

as desired.