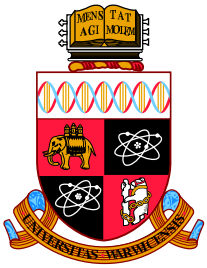
USING GENERALIZED LINEAR MODELS TO MODEL COMPOSITIONAL RESPONSE DATA

by
Fiona Sammut

A thesis presented for the degree of
Doctor of Philosophy



Department of Statistics
University of Warwick
September 2016

Dedication

To my husband, my constant

Contents

List of Figures

List of Tables

Acknowledgements

Without the help and support of a number of people, this work would not have been possible. In particular, I wish to thank:

My supervisor, Professor David Firth whose knowledge, guidance, stimulating suggestions and constructive criticism helped me throughout the different stages of preparing this thesis.

My friends Monique, Stephanie, David, Natalie and Shirley for their ever encouraging words and support.

Elena, Silvia, Panayiota and Pantelis for their friendship and for making me feel at home away from home.

My family whose thoughtful words made me a stronger person.

Especially, I would like to thank my husband, Trevor, for believing in me and supporting me through the good and the bad and whose patient and unconditional love enabled me to complete this work.

This research has been supported by the University of Malta; and by the Engineering and Physical Sciences Research Council [grant numbers EP/P503817/1, EP/P50578X/1, EP/J500586/1].

Declaration

I hereby certify that the work presented in this thesis is my own work and to the best of my knowledge, unless otherwise stated, this work is original and has not been submitted for a degree or a diploma to any other university.

Abstract

This work proposes a multivariate logit model which models the influence of explanatory variables on continuous compositional response variables. This multivariate logit model generalizes an elegant method that was suggested previously by Wedderburn (1974) for the analysis of leaf blotch data in the special case of $J = 2$, leading to our naming this new approach as the *generalized Wedderburn* method. In contrast to the logratio modeling approach devised by Aitchison (1982, J. Roy Stat. Soc. B.), the multivariate logit model used under the generalized Wedderburn approach models the expectation of a compositional response variable directly and is also able to handle zeros in the data. The estimation of the parameters in the new model is carried out using the technique of generalized estimating equations (GEE). This technique relies on the specification of a working variance-covariance structure. A working variance-covariance structure which caters for the specific variability arising in compositional data is derived. The GEE estimator that is used to estimate the parameters of the multivariate logit model is shown to be invariant to the values of the correlation and dispersion parameters in the working variance-covariance structure. Due to this invariance property and the fact that the estimating equations used under the generalized Wedderburn method are linear and unbiased, the GEE estimator achieves full efficiency across a wide class of potential dispersion and correlation matrices for the compositional response variables. As for any other GEE estimator, the estimator used in the generalized Wedderburn method is also asymptotically unbiased and consistent, provided that the marginal mean model specification is correct. The theoretical results derived in this thesis are substantiated by simulation experiments, and properties of the new model are also studied empirically on some classic datasets from the literature.

List of Notation

$\dot{\mathbf{Y}}$	vector of latent variables
$\mathbf{Y}$	composition
$\mathbf{W}$	vector of logratios
J	number of components in $\mathbf{Y}$ and $\dot{\mathbf{Y}}$
Y_j	j^{th} component in $\mathbf{Y}$
$\mathbb{X}$	design matrix
$\mathbb{X}_i$	design matrix corresponding to case i
X_k	k^{th} explanatory variable
S^{J-1}	$(J-1)$-dimensional unit simplex
$\mathbb{R}^J$	J-dimensional Euclidean space
$\mathbb{R}_+^J$	the positive orthant of J-dimensional Euclidean space $\mathbb{R}^J$
$C(\cdot)$	closure operation
$\boldsymbol{\beta}$	vector of model coefficients
$\boldsymbol{\gamma}$	identifiable vector of model coefficients
θ_i, $\dot{\theta}_i$	nuisance parameters for case i
$m_i(\theta, \boldsymbol{\beta})$	a function defined by $\exp\left(\theta + \mathbf{x}_i^{'}\boldsymbol{\beta}\right)$
p_{ij}	proportion for case i and compositional variable j
$\mathbf{p}_i$	vector of proportions $(p_{i1}, \ldots, p_{iJ})$ for case i
$\mathbb{P}_i$	diagonal matrix with proportions $(p_{i1}, \ldots, p_{iJ})$ for case i
ϕ	dispersion parameter
ω_j	relative dispersion parameter corresponding to component j
$\boldsymbol{\Omega}$	diagonal matrix of relative dispersion parameters
$\alpha_{jj'}$	the correlation between $\dot{Y}_{ij}$ and $\dot{Y}_{ij'}$
$\mathbb{W}$	working correlation matrix
$\mathbf{U}$	vector of estimating equations
$\dot{\mathbb{D}}$	matrix of derivatives of $m_i\left(\dot{\theta}_i, \boldsymbol{\beta}_j\right)$ with respect to $\boldsymbol{\beta}$ and $\boldsymbol{\theta}$
$\mathbf{1}$	vector of ones
$\mathbf{1}\mathbf{1}^{'}$	matrix of ones
$\mathbb{I}$	identity matrix
$\mathbb{V}_{\mathbf{p}_i,\boldsymbol{\Omega},\mathbb{W}}$	working variance-covariance matrix used under the generalized Wedderburn method
$\mathbb{V}_i$	implied form of the variance-covariance matrix of $\mathbf{Y}_i$
$\mathbb{A} \otimes \mathbb{B}$	the kronecker product of matrix $\mathbb{A}$ with matrix $\mathbb{B}$
$\mathbb{F}$	matrix of contrasts
$\boldsymbol{\Psi}$	variance-covariance matrix of $\log\left(\dot{\mathbf{Y}}\right)$
R_{ij}	Pearson residuals under the generalized Wedderburn method
R_{ij}^*	Aitchison residuals

List of Abbreviations

GEE	Generalized Estimating Equations
GLM	Generalized Linear Model
GLS	Generalized Least Squares
GW	Generalized Wedderburn
ML	Maximum Likelihood
MLE	Maximum Likelihood Estimator
MRM	Multiplicative Regression Model
OLS	Ordinary Least Squares

Chapter 1

Introduction

This thesis concerns regression modeling of compositional data, that is, models and associated methods used to describe the dependence of compositional responses upon explanatory variables. This chapter aims to provide the setting required, to better appreciate the challenges that have been faced in trying to find a suitable approach to model continuous compositional data. An introductory note on what compositional data is, is provided in Section 1.1. Section 1.2 then gives some details on a number of distributions that have been considered apt to model this type of data, namely the Dirichlet distribution and some of its generalizations, and the logistic normal distribution. The normal distribution, Barndorff-Nielsen and Jørgensen (1991) simplex distribution, the von Mises distribution and the Kent distribution (Kent, 1982) are also mentioned briefly. The logratio transformation (Aitchison, 1982) is presented in relation to the logistic normal distribution. Through this transformation it is possible to model the influence of explanatory variables on the transformed variables, using standard multivariate techniques. The main drawback with the logratio transformation is that it fails when any zeros are present in the data. A brief account of various ways in which researchers have tried to tackle the problem of zeros is thus provided in Section 1.3. Section 1.4 then describes some previous attempts at modeling compositional response data through the generalized linear modeling framework. Studies which used the method of generalized estimating equations (GEE) to model compositional response data are presented in Section 1.5. At the end of Section 1.5, a new approach which may be used to model continuous compositional response variables and which will be developed in this thesis, is introduced. This approach is also based on the technique of generalized estimating equations.

1.1 What is Compositional Data?

A composition $\mathbf{Y}$ is a J-vector whose components $Y_1, \ldots, Y_J$ are non-negative and satisfy a sum constraint, that is, $Y_1 + \cdots + Y_J = k$, where k is some fixed known constant that depends on unit of measurement. The constant is often equal to 1 or 100. Some examples

of how a composition may arise include measurements on the mineral and chemical content of rocks, household expenditure patterns, or time allocated to various activities during a particular day.

Due to the nature of compositional data, one part of the composition may always be written in terms of the remaining parts, effectively reducing the dimension of a J-part composition to $J-1$. The importance thus lies not with the actual value of a part in a composition but with the magnitude of a part in relation to the magnitude of the other parts in a composition. In much of the theoretical development of compositional data, the magnitude of the constant k in the sum-constraint, is set equal to 1. Fixing k at 1 gives rise to a $(J-1)$-dimensional unit simplex, S^{J-1}, embedded in a J-dimensional non-negative space. Without loss of generality, the material presented in this work will take k to be equal to 1.

Let $\dot{\mathbf{Y}}$ be a J-vector in the positive space $\mathbb{R}_+^J$, defined as $\dot{\mathbf{Y}} = \left(\dot{Y}_1, \ldots, \dot{Y}_J\right)'$, where each $\dot{Y}_j$ $(j = 1, \ldots, J)$, is measured in the same units and each $\dot{Y}_j$ provides relative information. The term relative refers to the fact that meaning to the information provided is given by the ratios of the different components. Due to only relative information being provided by the data, the unit of measurement chosen for $\dot{\mathbf{Y}}$ will make no difference to the analysis. In fact, any two such vectors, say $\dot{\mathbf{Y}}$ and $\dot{\mathbf{Y}}^*$, which are related by the equation $\dot{\mathbf{Y}}^* = a\dot{\mathbf{Y}}$, for some positive constant a, are regarded as equivalent. The equivalent vectors $\dot{\mathbf{Y}}^*$ and $\dot{\mathbf{Y}}$ may both be said to fall into the class $cl\left(\dot{\mathbf{Y}}\right) = \left\{a\dot{\mathbf{Y}} : a > 0\right\}$, the latter geometrically represented by a ray from the origin in $\mathbb{R}_+^J$. Intersecting this ray and the unit simplex S^{J-1} results in any vector in the class $cl\left(\dot{\mathbf{Y}}\right)$ to be sum-constrained to 1. This intersection, a constraining operation known as *closure*, relates the composition $\mathbf{Y}$ to any vector $\dot{\mathbf{Y}}$ in class $cl\left(\dot{\mathbf{Y}}\right)$ as follows

$$\mathbf{Y} = C\left(\dot{\mathbf{Y}}\right) = \frac{\dot{\mathbf{Y}}}{\dot{Y}_1 + \cdots + \dot{Y}_J}. \tag{1.1}$$

The fact that a whole class of vectors in $\mathbb{R}_+^J$ have the same closure $\mathbf{Y}$ leads to problems in analyzing compositional data using standard multivariate techniques directly. In particular, as per Aitchison (1986), the independence of the components of $\dot{\mathbf{Y}}$ would not correspond to any simple 'null' structure of the covariance between the components of the composition $\mathbf{Y}$. Restriction on the variance-covariance structure of the variables forming the composition arises in the fact that at least one covariance is forced, through the sum constraint, to be negative. As an example of the latter, consider $J = 2$ and the sum constraint $Y_1 + Y_2 = 1$. From this it follows that $\mathrm{Cov}\,(Y_1, Y_1 + Y_2) = 0$ and hence $\mathrm{Var}\,(Y_1) = -\mathrm{Cov}\,(Y_1, Y_2)$ automatically. This restriction on $\mathrm{Cov}\,(Y_1, Y_2)$ will then lead to a restriction in the correlation coefficient.

Pearson (1897) had already put forward the possibility of obtaining a 'high degree of correlation between series of absolutely independent judgements' when dealing with com-

positional data. Nothwithstanding this fact, scientists and statisticans alike still resorted to the standard multivariate techniques to analyze this type of data and it was only around the 1960s that papers which expressed disapproval towards using the usual statistical approach started to emerge.

The most prominent critic of the time was the geologist Felix Chayes. His criticism was mainly focused on the interpretation of the correlation coefficient between components in a geochemical composition. Chayes (1960) stated 'neither the resulting spurious correlation itself nor the difficulty it creates with regard to the interpretation of composition data has been adequately described, and no general remedy has yet been suggested'.

A lot of effort has been directed in trying to obtain a formulation for this spurious correlation, or mostly known as *null correlation* as opposed to the zero correlation resulting from lack of dependence in the usual statistical sense. Papers by Darroch (1969), Meisch (1969), Darroch and Ratcliff (1970), Darroch and Ratcliff (1978) and Kork (1977), all dealt with the issue of the sum-to-a-constant constraint and its relation to the interpretation of correlations of proportions. Aitchison (1984) still felt the need to encourage researchers, petrologists in particular, to apply the right methodology in dealing with compositional data. Illustrations and warnings on why standard statistical techniques would fail had been issued back then, but were largely disregarded.

According to Aitchison (1982, 1984a), it was the 'lack of concepts of independence, lack of a satisfactory and interpretable covariance structure and the lack of parametric classes of distributions' which were appropriate for the simplex, that hindered the development of suitable methods for such data. To this end, Aitchison (1986) and Pawlowsky-Glahn and Egozcue (2006) recognized the fact that until the early 1980s there was no clear guidance on how to deal with compositional data.

1.2 In Search of a Family of Suitable Distributions

1.2.1 The Dirichlet Distribution

Prior to the 1980s, the only familiar class of distributions which was thought to be appropriate for the space of all continuous compositions was the Dirichlet family, which is defined as follows:

Definition 1.2.1. *Let* $\mathbf{Y} = (Y_1, \ldots, Y_J)'$ *where* $Y_J = 1 - \sum_{j=1}^{J-1} Y_j$ *and let* $\boldsymbol{\alpha} = (\alpha_1, \ldots, \alpha_J)'$ $\in \mathbb{R}_+^J$. $\mathbf{Y}$ *is said to follow the Dirichlet distribution on* S^{J-1} *with parameter* $\boldsymbol{\alpha}$ *if its probability density function is given by*

$$f(y_1, \ldots, y_{J-1}|\boldsymbol{\alpha}) = \frac{\Gamma(\alpha_1 + \cdots + \alpha_J)}{\prod_{j=1}^{J} \Gamma(\alpha_j)} \prod_{j=1}^{J} y_j^{\alpha_j - 1}, \tag{1.2}$$

where $(y_1, \ldots, y_J)'$ *denotes a vector of values of* $(Y_1, \ldots, Y_J)'$ *and* $y_j \in (0, 1)$.

Letting $\alpha_+ = \alpha_1 + \cdots + \alpha_J$, some well known properties of the Dirichlet distribution are

$$E(Y_j) = \frac{\alpha_j}{\alpha_+}, \tag{1.3}$$

$$\text{Var}(Y_j) = \frac{\alpha_j(\alpha_+ - \alpha_j)}{\alpha_+^2(\alpha_+ + 1)}, \tag{1.4}$$

and for $j \neq j^{'}$

$$\text{Cov}\left(Y_j, Y_{j'}\right) = \frac{-\alpha_j \alpha_{j'}}{\alpha_+^2(\alpha_+ + 1)} \tag{1.5}$$

$$\text{Corr}\left(Y_j, Y_{j'}\right) = -\sqrt{\frac{\alpha_j \alpha_{j'}}{(\alpha_+ - \alpha_j)\left(\alpha_+ - \alpha_{j'}\right)}}. \tag{1.6}$$

Gueorguieva et al. (2008) and Maier (2014) are examples of researchers that have used Dirichlet regression models to model compositional response variables. Focusing on (1.5) and (1.6), it may however be noted that the correlation structure of the Dirichlet distribution is completely negative. The Dirichlet distribution would thus not be able to cater for non-negative correlations in the data, should any be present.

Also, the Dirichlet distribution inherits an independence structure through its own definition. Correlations between components $(Y_1, \ldots, Y_J)$ arise solely through the closure operation. As per Aitchison (1982), if a composition $\mathbf{Y}$ is assumed to follow a Dirichlet distribution, then $\mathbf{Y}$ may be considered to be the result of a closure operation performed on independent gamma distributed random variables, each having the same scale parameter.

Particularly due to this independence structure, Connor and Mosimann (1969), Darroch and James (1978), James and Mosimann (1980), Barndorff-Nielsen and Jørgensen (1991), Rayens and Srinivasan (1994), Gupta and Richards (1987, 1991, 1992, 1995), Aitchison (2003b, p. 305-306), amongst others, attempted to find generalizations of the Dirichlet class of distributions which could also accomodate the dependence between variables. As per Aitchison (2003b, p. 305), finding a 'tractable parametric class which contains the Dirichlet distribution but which also caters for significant departure from its strong independence properties' is still considered an open problem. The flexible Dirichlet distribution (Ongaro and Migliorati, 2013), 'obtained by normalizing a correlated basis formed by a mixture of independent gamma random variables' and of which the Dirichlet distribution is a special case, is promising in the fact that it allows for a more flexible dependence structure whilst retaining the same mathematical and compositional properties of the Dirichlet distribution. (For more detail on these properties refer to Ongaro and Migliorati (2013).) As per Migliorati et al. (2016), the flexible Dirichlet 'displays a dependence structure that is substantially richer' than the generalized Dirichlet distribution (Connor and Mosimann, 1969), the simplex distribution (Barndorff-Nielsen and Jørgensen, 1991) and the multivariate Liouville distribution (Gupta and Richards, 1987, 1991, 1992, 1995) and 'a greater

tractability than the generalized Liouville distribution' (Rayens and Srinivasan, 1994). Migliorati et al. (2016) also state that through its mixture structure, the flexible Dirichlet distribution may be used to model various features of a compositional data set, 'including unimodal and multimodal cases'. However Migliorati et al. (2016) recommend to develop 'more flexible structured mixture models' which are 'still inferentially tractable'.

1.2.2 The Logistic Normal Distribution

As an alternative to modeling the compositional data directly in the simplex, a parametric class of distributions which could cater for the dependence structure between the parts of the compositions but which could also make the transition from the the positive real line to the whole real line possible was devised. McAlister (1879) realized that if he considered a normally distributed random variable, by taking its exponent, a useful distribution (the lognormal) would be induced on the positive real line. Throughout the 20th century, especially following the work on variance-stabilizing transformations for analysis of variance, the general Box-Cox transformation (Box and Cox, 1964) and other work on transformations to normality continued to emerge. In spite of this, it was only in the early 1980s that a new method based on McAlister's (1879) idea of inducing a distribution on an 'awkward' space, from another distribution defined on a more familiar space, by using a transformation between the two spaces, was devised.

Prior to the 1980s, the logistic-normal distribution had already been used in areas like Bayesian analysis for the description of a prior and posterior distribution of multinomial probabilities (Lindley, 1964), logistic disciminant analysis (Anderson, 1972) and in analyzing binary data. The application of the logistic-normal distribution to the field of compositional data is attributed to Aitchison and Shen (1980). As per Aitchison and Shen (1980), if the logistic transformation is applied to a $(J-1)$-vector $\mathbf{W} \in \mathbb{R}^{J-1}$, where $\mathbf{W}$ follows the multivariate normal distribution, the resulting vector may be said to follow a logistic-normal distribution.

Definition 1.2.2. *For $\mathbf{Y} \in S^{J-1}$ and $\mathbf{W} \in \mathbb{R}^{J-1}$, the generalized logistic transformation, also known as the additive logistic transformation, is given by*

$$Y_j = \begin{cases} \frac{\exp(W_j)}{1+\sum_{j'=1}^{J-1}\exp\left(W_{j'}\right)} & \text{if } j = 1, \ldots, J-1 \\ \frac{1}{1+\sum_{j'=1}^{J-1}\exp\left(W_{j'}\right)} & \text{if } j = J, \end{cases}$$

where for $j = 1, \ldots, J-1$, its inverse function is given by

$$W_j = \log\left(\frac{Y_j}{Y_J}\right). \tag{1.7}$$

The inverse function (1.7) is known as the additive logratio (alr) transformation.

So as to overcome the problem of analyzing compositional data, Aitchison (1982) proposed the use of logratio transformations, amongst which is the additive logratio transformation. If the multivariate normal distribution may be shown to be a reasonable approximation to the distribution of the logratios $W_1, \ldots, W_{J-1}$, conventional multivariate techniques may be used on the logratio-transformed data and standard statistical analysis may then be carried out in the real space $\mathbb{R}^{J-1}$. Aitchison (1982) used this idea to model the influence of explanatory variables $X_1, \ldots, X_p$ on compositional response variables: for a sample of size n and J components,

$$E\left[\log\left(\frac{Y_{ij}}{Y_{iJ}}\right)\right] = \mathbf{x}_i^{'}\boldsymbol{\beta}_j \tag{1.8}$$

where $i = 1, \ldots, n$, $j = 1, \ldots, J-1$, $\boldsymbol{\beta}_j$ is a $(p+1)$-vector of coefficients which need to be estimated and $\mathbf{x}_i$ is a $(p+1)$-vector of observations with $x_{i0} = 1$ since the first element corresponds to the intercept and the remaining elements correspond to the observations obtained by the ith case in the sample on $X_1, \ldots, X_p$.

The method of modeling the conditional expectation of logratios is appealing for a number of reasons, including the fact that it is permutation invariant. So if any component other than Y_J is chosen as the reference component in the additive logratio, the same results will be obtained (Aitchison, 1986, p. 96). As per Aitchison (1986, p. 96), standard multivariate statistical procedures are all invariant under the group of permutations of the parts $1, \ldots, J$ of the composition.

Statisticians seemed to accept such methodology. However, transformation resistance syndrome, as described by Aitchison (2003a), especially amongst the geological community (refer to letters to the Editor of Mathematical Geology over the period 1988 to 2002; Rehder and Zier (2001); Aitchison and Barceló-Vidal (2001)), still prevailed. Some of the arguments brought up dealt with the theoretical nature of the subject. Others dealt with the difficulty in interpreting the transformed data.

The additive logratio transformation, in particular, does in fact present some problems related to interpretation. It is asymmetric in the parts of the composition and by changing the denominator, a different transformation results and thus also a different interpretation has to be given. One major shortcoming that is common to all logratio transformations is that they may only be applied to compositions whose parts are strictly positive. If some parts of a composition are zero, the corresponding logratios cannot be computed. More details on methods that were developed to deal with zeros in compositional data are provided in Section 1.3.

1.2.3 Other Distributions

1.2.3.1 The Normal Distribution

After searching the literature using key words such as 'percentage, proportion and fraction', Kieschnick and McCullough (2003) found that the majority of researchers used the normal distribution as the conditional distribution of 2-part compositional data given a set of explanatory variables. Since compositional data is not defined over the whole real line, though, it can never follow a normal distribution.

Other attempts at using the normal distribution to analyze compositional data have also been made through the use of latent variable models. Some detail and criticism on a censored normal model and on the Butler and Glasbey (2008) latent Gaussian model will be given in Section 1.3.2.3.

1.2.3.2 Barndorff-Nielsen and Jørgensen (1991) Simplex Distribution

Another distribution which has been used to model 2-part compositional data is the univariate simplex distribution (Barndorff-Nielsen and Jørgensen, 1991). The density function of a response variable Y which follows a univariate simplex distribution is given by

$$f\left(y|\mu,\sigma^2\right) = \frac{1}{\sqrt{2p\sigma^2\left[y\left(1-y\right)^3\right]}} \exp\left[-\frac{1}{2\sigma^2} d\left(y;\mu\right)\right],$$

for $0 < y < 1$, $0 < \mu < 1$ and where $d\left(y;\mu\right) = \left(y-\mu\right)^2 / y\left(1-y\right)\mu^2\left(1-\mu\right)^2$ is the unit deviance. The form of the density of the univariate simplex distribution implies that the simplex distribution is a proper dispersion model (see Jørgensen, 1986) where the parameters μ and σ^2 correspond to the position and dispersion parameters. This distribution is however defined on the open interval $(0, 1)$, excluding the possibility of analyzing compositional data with zeros.

In Section 1.5, we will see how the Barndorff-Nielsen and Jørgensen (1991) simplex distribution has been used in a generalized linear modeling setup in attempt to model compositional data. Zhang (2013) made use of a multivariate version of this distribution, which is however also not suitable to model zeros in compositional data. More details on the use of the multivariate simplex distribution Barndorff-Nielsen and Jørgensen (1991) in Zhang (2013) will also be given in Section 1.5.

1.2.3.3 Distributions Defined on the Hypersphere

The Kent distribution and the von Mises-Fisher distribution (for more information on these distributions refer to (Mardia and Jupp, 2000)) are another two distributions which have

been used to model compositional data. More specifically, the two distributions have been used by Scealy and Welsh (2011) and Stephens (1982) respectively, to model data defined on the hypersphere. In directional data, a direction in J dimensions may be represented by a vector, say $\mathbf{S}$, which lies on the surface of a $(J-1)$-dimensional hypersphere of unit radius and is centred at the origin. Vectors on the surface of the hypersphere satisfy $\mathbf{S}'\mathbf{S} = \mathbf{1}$. Compositional data is turned into directional data by means of the square root transformation, that is by letting $\mathbf{S} = \left(\sqrt{Y_1}, \ldots, \sqrt{Y_J}\right)$. By the sum-to-1 constraint of compositional data, $\mathbf{S}'\mathbf{S} = \mathbf{1}$ follows naturally.

Stephens (1982) used the von Mises-Fisher distribution to discriminate between two groups of students. Aitchison (2008) acknowledges the fact that a 'reasonable discrimination' has been achieved on using such an approach but discourages further use of such an approach due to the simplex and the sphere being 'topologically completely unrelated'. Mardia (1976) also showed that the von Mises-Fisher distribution can provide a good approximation to the multinomial distribution, showing that the von Mises-Fisher distribution inherits the restrictive correlation structure of the multinomial distribution, making the von Mises-Fisher distribution unfit to model compositional data.

As per Scealy and Welsh (2011), the Kent distribution on the other hand, provides a 'natural generalization' on the von Mises-Fisher distribution, has as many parameters as the $(J-1)$-variate normal distribution, and its parameters are 'readily interpretable'. Some more detail on how Scealy and Welsh (2011) have successfully managed to use the Kent distribution to model the influence of explanatory variables on compositional response variables will be given in Section 1.3.2.4.

1.3 Analyzing Compositional Data with Zeros

Compositions with zero values may easily turn up in practice. An example which is very frequently used in the literature is that of household expenditure where some families would spend nothing on alcohol and cigarettes. Zeros might also be obtained in an analysis of time allocated to different tasks with some people allocating no time to physical activity. Furthermore, zeros might also result in an analysis of the proportion of fat, carbohydrates and protein in food items, with no protein or fat found in sugar. It might also be of interest to study the presence of various species in an area with some species remaining undetected.

1.3.1 Rounded and Essential Zeros

When it comes to dealing with zeros in the data, a distinction needs to be made with respect to the type of zero that may occur. If accuracy limitations of instruments of measurements or some other physical, chemical or artificial effects prevent us from detecting small concentrations of some part or parts of a composition, these small concentrations

may be considered as having been erroneously recorded as zeros (Palarea-Albaladejo et al., 2007). When this type of zero arises it is referred to as a *rounded* zero. Aitchison (1986, p. 268) proposed to investigate ways in which rounded zeros in compositional variables may be replaced by some small positive value which is smaller than the smallest recordable value. Various imputation methods for rounded zeros, amongst which are the multiplicative replacement technique (Martín-Fernández et al., 2003) and the modified EM algorithm (Palarea-Albaladejo and Martín-Fernández, 2008) exist in the literature. So once the zeros in the data have been imputed, a logratio transformation may then be applied and the transformed data is then analysed using standard statistical techniques.

Imputation may not however be used when there are *essential* zeros in the data. An essential zero in a composition is a zero which may not be considered to be the result of a limitation of the measuring instrument being used but is the result of something that is completely absent. An essential zero might be obtained, for example, from a family whose expediture on household appliances during a study period is nil. In some cases, it might make sense to overcome modeling problems due to essential zeros by performing a separate analysis of subjects/objects coming from different groups/subpopulations. Alternatively, it might also be reasonable to amalgamate the proportions from different components of the composition. After amalgamation of the different parts, statistical analysis may then be carried out on the full dataset. Models that deal with the essential zero problem have been proposed but at this stage there is still no standard procedure which should be implemented.

1.3.2 Various Attempts at Solving the Essential Zero Problem

1.3.2.1 The Addition of a Constant to Every Observation

The first attempt at solving the essential zero problem came from Aitchison (1986). Aitchison (1986, p. 271) proposed to take insight from the three-parameter lognormal model (Aitchison and Brown, 1957) where a constant υ_j, $j = 1, \ldots, J$, known or to be estimated is added to every observation and the logratio transformation is then applied to the vector $C(\mathbf{Y} + \boldsymbol{\upsilon})$ rather than to $C(\mathbf{Y})$. Clearly, for a dataset with a large number of zeros, this technique will lead to substantial computational effort. Furthermore, even with compositions with a relatively small number of parts, with the inclusion of υ_j, a serious interpretation problem of the transformed data arises.

1.3.2.2 Conditional Modeling

Aitchison (1986) also suggested the alternative idea of using some form of conditional modeling to deal with the essential zero problem. Aitchison (1986, p. 272) provides an example where zero values arise in the first component only. A probability p is assigned to the probability of Y_1 being equal to zero. The probability of Y_1 not being a zero is

then $(1-p)$. Conditional on $Y_1 = 0$, the distribution for $\mathbf{Y}_{(1)} = (Y_2, \ldots, Y_J)'$ is taken to be the additive logistic normal distribution with parameters $\boldsymbol{\mu}_0$ and $\boldsymbol{\Sigma}_0$. Conditional on $Y_1 > 0$, $\mathbf{Y}$ is taken to follow an additive logistic normal distribution with parameters $\boldsymbol{\mu}$ and $\boldsymbol{\Sigma}$. Let $f_{J-1}(\cdot|\boldsymbol{\mu}, \boldsymbol{\Sigma})$ denote the probability density function corresponding to $\mathbf{Y}$. The contribution to the likelihood by a compositional vector with $Y_1 = 0$ is then

$$p f_{J-2}(\cdot|\boldsymbol{\mu}_0, \boldsymbol{\Sigma}_0)$$

and the contribution to the likelihood by a compositional vector with $Y_1 > 0$ is given by

$$(1-p) f_{J-1}(\cdot|\boldsymbol{\mu}, \boldsymbol{\Sigma}).$$

Estimation of the parameters may then be carried out through maximum likelihood estimation.

The setup provided by Aitchison (1986) is however a very simple one. Bacon-Shone (2003) explains that a suitable model is one which is able to handle two problems; the first being that of modeling the pattern of zeros for multiple components and the second being that of modeling a composition conditional on the particular pattern of zeros which arises. Aitchison and Kay (2003), Zadora et al. (2010) and Tsagris (2014) also propose to use a two-step approach to model compositional data with zeros. Zadora et al. (2010) and Tsagris (2014) first model the presence of zeros using the independent binary model

$$\prod_{j=1}^{J} p_j^{u_j} (1-p_j)^{1-u_j},$$

where p_j is the probability of obtaining a non-zero value in the j^{th} component and u_j is an indicator function taking the value 1 if the j^{th} component in a composition is not zero and a value of 0 in the presence of a zero. A multivariate density function which incorporates the model for the zeros in then adopted in the second stage.

In relevance to the nature of compositional data, the assumption of independence in the independence binary model is completely violated. Aitchison and Kay (2003) thus impose a hierarchical prior on the binomial parameters p_j,

$$p_j = \frac{\exp(\lambda_j)}{\exp(\lambda_j) + 1},$$

where the parameters $(\lambda_1, \ldots, \lambda_J)$ may be assumed to follow a multivariate normal distribution. As per Aitchison and Kay (2003), the issues related with making use of this strategy are more of the computational kind with the main problem being the evaluation of the integral in the dependent binomial case. Pertaining to the latter problem, Aitchison and Kay (2003) expected the MCMC approach to be conducive.

1.3.2.3 The Latent Model

Alternative attempts at modeling compositional data with zeros involved the use of latent variable models. The Tobit model and Butler and Glasbey (2008) latent Gaussian model are two such examples. Both models may be used in the presence of zeros. Both of them, however, are not suitable to model compositional data. More detail about the two models follows.

The Tobit Model

There are occasions (e.g. Agnew et al., 1995; Barclay and Smith, 1995) where the censored normal model, also known as the Tobit model, has been used to model continuous proportions. Continuous proportions may be viewed as 2-part compositional data. For a dataset of size n, in a Tobit model, a latent (unobservable) variable $\dot{Y}$ is related to p independent variables $X_1, \ldots, X_p$, through the linear model

$$\dot{Y}_i = \mathbf{x}_i' \boldsymbol{\beta} + E_i, \quad i = 1, \ldots, n$$

and the response variable Y satisfies

$$Y_i = \begin{cases} 0 & \text{if } \dot{Y}_i \leq 0 \\ \dot{Y}_i & \text{if } 0 < \dot{Y}_i < 1 \\ 1 & \text{if } \dot{Y}_i \geq 1 \end{cases}$$

where $\boldsymbol{\beta}$ is a vector of unknown regression parameters and E_i are error terms which are assumed to be independent and identically $N\left(0, \sigma^2\right)$ distributed.

This model may handle zeros in the data but values outside the range $(0, 1)$ are being treated as if they were censored. It is due to the nature of compositional data that values beyond the range $[0, 1]$ may never be observed.

Butler and Glasbey (2008) Latent Gaussian Model

Butler and Glasbey (2008) propose to handle zeros in compositional data by means of a latent Gaussian model. This model is based on the assumption that compositional data is obtained through performing a transformation on an underlying set of latent variables $\dot{\mathbf{Y}}$. These latent variables are assumed to follow a multivariate Gaussian distribution with unknown parameters $\boldsymbol{\mu}$ and $\boldsymbol{\Sigma}$, and $\dot{\mathbf{Y}}$ is also assumed to lie on the unit hyperplane $H_J = \left\{\dot{\mathbf{y}} \in \mathbb{R}^J : \dot{\mathbf{y}}' \mathbf{1} = 1\right\}$. Such an assumption on $\dot{\mathbf{Y}}$ implies that the mean vector $\boldsymbol{\mu}$ and the variance-covariance matrix $\boldsymbol{\Sigma}$ of $\dot{\mathbf{Y}}$ must satisfy $\boldsymbol{\mu}'\mathbf{1} = 1$ and $\boldsymbol{\Sigma}\mathbf{1}\mathbf{1}' = \mathbb{O}$, where $\mathbf{1}\mathbf{1}'$ is a $J \times J$ matrix of ones. The chosen function which transforms $\dot{\mathbf{Y}}$ is taken to be that which minimizes the squared Euclidean distance between $\dot{\mathbf{Y}}$ and $\mathbf{Y}$, subject to the

usual sum constraint and non-negativity of the parts of the compositions. The function performs a Euclidean projection from a unit hyperplane onto a unit simplex. Butler and Glasbey (2008) explain that the choice of the Euclidean transformation was motivated by its wide use and its pleasing theoretical properties.

The assumption that the latent variables following a multivariate normal distribution makes the estimation process relatively easy for $J \leq 3$ but the authors acknowledge that parameter estimation becomes complicated once $J > 3$. Also, by assuming that the compositional data comes into existence through the multivariate normal distribution, the principles of scale invariance and subcompositional invariance are violated. The principle of scale invariance requires that any function of compositional data should be expressed in terms of ratios and the principle of subcompositional invariance requires that the same information is obtained from the components which are common to different subsets of a composition. Butler and Glasbey (2008) thus recommend that this approach is only used if i) other methods are not available or if they are inappropriate, ii) as a diagnostic tool for assessment of other methods or iii) for exploratory purposes. Butler and Glasbey (2008) also suggest to investigate the possibility of using alternative transformations.

1.3.2.4 The Square-Root Transformation

Scealy and Welsh (2011) propose an alternative approach to handle zeros in compositional data. A brief mention of this approach has also been given in Section 1.2.3.3. It is based on applying the square root transformation on compositional data, including zeros in the data, so that compositional data is transformed to directional data. Stephens (1982) and Scealy and Welsh (2011) used the square root transformation as a first step in their analysis to model compositional data. Scealy and Welsh (2011) proceed to develop a regression model for compositional data through the Kent distribution (Kent, 1982), by relating the mean direction vector to linear functions of the explanatory variables. The performance of the estimation procedure described in Scealy and Welsh (2011) is however impaired if the majority of the transformed data is close to the boundaries of the positive orthant. So Scealy and Welsh (2014) revise the estimation procedure described in Scealy and Welsh (2011) and also show how the EM algorithm may be used to estimate the parameters of the folded Kent distribution so as to deal with the problem of having a large concentration of points close to the boundaries with a 'relatively large variance'. The Kent regression model as a means to model compositional data with zeros is promising but it lacks simplicity of implementation.

1.3.2.5 The α-Transformation

Tsagris et al. (2011) propose the use of the α-transformation, a Box-Cox type of transformation, which allows more flexibility in analyzing compositional data than the logratio approach. The α-transformation is based on the power transformation proposed by

(Aitchison, 1986, p. 120).

For a composition $\mathbf{Y}$ and any real number α, the power transformation proposed by Aitchison (1986) is given by

$$\mathbf{S} = \left(\frac{Y_1^\alpha}{\sum_{j=1}^J Y_j^\alpha}, \ldots, \frac{Y_J^\alpha}{\sum_{j=1}^J Y_j^\alpha} \right)'. \tag{1.9}$$

The α-transformation is then defined by

$$\mathbf{T} = \frac{1}{\alpha} \mathbb{H} \left(J\mathbf{S} - \mathbf{1} \right), \tag{1.10}$$

where $\mathbf{1}$ is a J-dimensional vector of ones and $\mathbb{H}$ is the $(J-1) \times J$ Helmert submatrix (Lancaster, 1965), the latter 'obtained by removing the first row from the Helmert matrix' (Tsagris et al., 2011).

Tsagris (2015) proposes an approach involving the α-transformation to model compositional response variables, even in the presence of zeros. Tsagris (2015) calls this method α-regression.

The steps in α-regression start by assuming that, for each case i, the conditional expectation of the compositional response variables are given by

$$\mu_{ij} = \begin{cases} \frac{1}{1+\sum_{j'=1}^J \exp\left(\mathbf{x}_i' \boldsymbol{\beta}_{j'}\right)} & \text{for } j = 1 \\ \frac{\exp\left(\mathbf{x}_i' \boldsymbol{\beta}_j\right)}{1+\sum_{j'=1}^J \exp\left(\mathbf{x}_i' \boldsymbol{\beta}_{j'}\right)} & \text{for } j = 2, \ldots, J. \end{cases}$$

The use of the multinomial logistic function ensures that, for each i, $\sum_{j=1}^J \mu_{ij} = 1$.

The conditional means and the compositional response variables for each i are then α-transformed to get

$$\mathbf{Y}_{i\alpha} = \frac{1}{\alpha} \mathbb{H} \left(J \frac{Y_{i1}^\alpha}{\sum_{j=1}^J Y_{ij}^\alpha} - 1, \ldots, J \frac{Y_{iJ}^\alpha}{\sum_{j=1}^J Y_{ij}^\alpha} - 1 \right)'$$

and

$$\boldsymbol{\mu}_{i\alpha} = \frac{1}{\alpha} \mathbb{H} \left(J \frac{\mu_{i1}^\alpha}{\sum_{j=1}^J \mu_{ij}^\alpha} - 1, \ldots, J \frac{\mu_{iJ}^\alpha}{\sum_{j=1}^J \mu_{ij}^\alpha} - 1 \right)'.$$

This is then followed by assuming that the α-transformed vector of response, $\mathbf{Y}_{i\alpha}$, follows a multivariate normal distribution with the mean vector $\boldsymbol{\mu}_{i\alpha}$. Multivariate regression is then used to model the whole vector of α-transformed responses $\mathbf{Y}_\alpha$. The chosen value of

α is that which minimizes twice the Kullback-Leibler divergence (Kullback, 1997), that is

$$\text{KL} = 2\sum_{i=1}^{n}\sum_{j=1}^{J} y_{ij}\log\left(\frac{y_{ij}}{\widehat{y}_{ij}}\right),$$

where y_{ij} is the i^{th} observation obtained on compositional Y_j 'and $\widehat{y}_{ij}$ is the corresponding fitted value' (Tsagris, 2015).

The results in Tsagris (2015) show that α-regression might lead to better prediction than when data is modeled using the standard logratio approach and this approach is also able to handle zeros in the data. As per Tsagris (2015), however, unless the chosen value of α is the true value, the estimator of the β parameters will not be consistent. Also, the choice of the multivariate normal distribution to model the transformed data might not be ideal since the α-transformation maps the data onto a subset of $\mathbb{R}^{J-1}$, whilst the multivariate normal distribution operates over the whole of $\mathbb{R}^{J-1}$.

1.4 A Generalized Linear Modeling (GLM) Framework for Compositional Data

Frequently, researchers examine the influence of selected variables on response variables through either modeling directly the conditional expectation of the response variables or otherwise modeling a function of the conditional expectation of the response variables. In the case of modeling compositional response variables, the response values may vary in the range [0,1] (vector of proportions/fractions/percentages). So if an analyst is interested in examining how explanatory variables $X_1, \ldots, X_p$ influence a compositional response vector $\mathbf{Y}$, the model used has to accommodate the relationship which arises between p predictor variables $X_1, \ldots, X_p$ and J response variables $Y_1, \ldots, Y_J$ such that the sum constraint $Y_1 + \cdots + Y_J = 1$ holds.

In Section 1.2.2 it has been mentioned how Aitchison (1982) developed the strategy of modeling compositional response variables by performing regression modeling on the logratios of the compositions (refer to equation (1.8)). The major drawback of this technique, however, is that it breaks down when a compositional response variable takes on an exact value of 0.

In general, two important aspects, adapted from Kieschnick and McCullough (2003), have to be taken note of when modeling a compositional response variable as a linear function of explanatory variables. These are:

1. the conditional expectation should be nonlinear since it maps onto the bounded interval [0,1]
2. the conditional variance should be a function of the mean since the variance will approach zero as the mean approaches either boundary point.

In this section, alternative ways of modeling compositional data, using a GLM framework, will be presented.

1.4.1 The Two-Parameter Beta Distribution in Conjunction with a Logit-Link Function

Based on the work of Cox (1996) who tested various link functions for regression models of continuous proportions, Kieschnick and McCullough (2003) consider a 2-part composition and use the logit link to specify the relationship between the conditional expectation of Y_1 and the vector of predictors $\mathbf{X}$ as follows:

$$E\left(Y_{i1}\right) = \frac{1}{1 + \exp\left(-\mathbf{x}_i'\boldsymbol{\beta}\right)}. \tag{1.11}$$

Kieschnick and McCullough (2003) then assume that the response variable Y_1 follows a two-parameter beta distribution[1]. Kieschnick and McCullough (2003) give two reasons for choosing this distribution for compositional data, the first being that it is the distribution most often fitted to fractional response variables in prior literature, the second being that the two-parameter beta distributions form an exponential family.

If Y_1 follows a two-parameter beta distribution, its probability density function is given by

$$f\left(y_1\right) = \frac{1}{B\left(p, q\right)} y_1^{p-1}\left(1 - y_1\right)^{q-1}, \tag{1.12}$$

where $0 \leq y_1 \leq 1$ and $B\left(p, q\right)$ is the beta function and its expected value is given by

$$E\left(Y_1\right) = \frac{p}{p + q}. \tag{1.13}$$

Hence, Kieschnick and McCullough (2003) relate equations (1.11) and (1.13) to obtain

$$q\left(\mathbf{x}_i\right) = p \exp\left(-\mathbf{x}_i'\boldsymbol{\beta}\right). \tag{1.14}$$

Substitution of (1.14) into (1.12) then yields the conditional density function

$$f\left(y_1|\mathbf{x}_i\right) = \frac{1}{B\left(p, q\left(\mathbf{x}_i\right)\right)} y_1^{p-1}\left(1 - y_1\right)^{q(\mathbf{x}_i)-1}, \tag{1.15}$$

from which the conditional expectation of Y_1 may be calculated once the parameters $\boldsymbol{\beta}$ and p are estimated through maximum likelihood estimation.

[1]The two-parameter beta distribution is a special case of the Dirichlet distribution. The Dirichlet distribution has been defined in Section 1.2.

This approach manages to restrict the conditional mean of a beta distributed random variable to the interval $(0, 1)$ and a multivariate generalization of it may be obtained by using the Dirichlet distribution instead. However, estimates of the conditional expectation obtained when the two-parameter beta distribution is used, are known not to be robust to distributional failure (Papke and Wooldridge, 1996). The reason behind this follows from the fact that the two-parameter beta distributions do not form a linear exponential family (neither do the Dirichlet distributions) and thus, consistency of the estimator is only guaranteed if the score equations from the beta log-likelihood are unbiased. This unbiasedness will only hold if the true generating process is beta but not otherwise (Gourieroux et al., 1984). Consequently, Kieschnick and McCullough (2003) and Papke and Wooldridge (1996) argue that a better approach to model fractional data is the quasi-likelihood approach developed by Wedderburn (1974).

1.4.2 Using a Quasi-Likelihood Approach

Only the first two moments need to be specified when using a quasi-likelihood approach and the variance has to be a function of the mean. Wedderburn (1974) presented the theoretical framework of quasi-likelihood estimation and used a logit link function together with a mean-variance relationship defined by $V(\mu_{i1}) = \mu_{i1}^2 (1 - \mu_{i1})^2$ to model the proportion of barley leaf area that was infected with *Rhynchosporium secalis* ('leaf blotch'), where $\mu_{i1} = E(Y_{i1})$. Despite the fact that the aim in Wedderburn (1974) was not directed towards tackling the problem of analyzing compositional data, the response variable analyzed in the paper may be viewed as arising out of a 2-part composition. The theoretical framework provided in this paper led the way to develop alternative ways of imposing structure on compositional data.

Papke and Wooldridge (1996), in fact, consider a 2-part composition, model the mean-variance relationship of Y_{i1} as for a Bernoulli distribution, and suppose that the explanatory variables influence the compositional response variable Y_{i1} through

$$E(Y_{i1}) = G\left(\mathbf{x}_i'\boldsymbol{\beta}\right), \tag{1.16}$$

where $G(\cdot)$ is a known function which satisfies $0 < G(\cdot) < 1$. This approach can handle any zeros that might be present in the data and also ensures that the predicted values of Y_1 given $\mathbf{x}_i$ lie in the interval $(0, 1)$. In fact, G is usually taken to be a cumulative distribution function such as the logistic function $G(z) = \frac{\exp(z)}{1+\exp(z)}$ or $G(z) = \boldsymbol{\Phi}(z)$ where $\boldsymbol{\Phi}(\cdot)$ is the cumulative distribution function for the standard normal distribution.

This approach is attractive for a number of reasons. On choosing the function $G(\cdot)$ such that it satisfies $0 < G(\cdot) < 1$, the log-likelihood function is well defined, may be generalized to cater for more than two response variables through the multinomial distribution, may be used even if zeros are in the data and it may be easily maximized using maximum

likelihood estimation. With respect to this, it might be argued that since a Bernoulli likelihood function is being used, this approach might simply be thought of in terms of maximum likelihood estimation rather than quasi-likelihood. It should however be emphasized that it is not actually the case that a fractional response variable follows a Bernoulli distribution. The observations making up Y_1 are being treated as pseudo-Bernoulli and since Bernoulli distributions form a linear exponential family, provided that (1.16) holds, the quasi-likelihood estimator will be a consistent estimator of the true vector of parameters $\boldsymbol{\beta}$ and it will also be $\sqrt{n}$-asymptotically normal, irrespective of the actual distribution of Y_1.

Now, if following the ideas of McCullagh and Nelder (1989) and Cox (1996), the logit link was to be used to specify the relationship between the conditional expectation of Y_1 and the vector of predictors $\mathbf{X}$, the conditional variance is given by

$$\text{Var}\left(Y_{i1}\right) = \sigma^2 G\left(\mathbf{x}_i'\boldsymbol{\beta}\right)\left(1 - G\left(\mathbf{x}_i'\boldsymbol{\beta}\right)\right) \tag{1.17}$$

for some variance $\sigma^2 > 0$. In case of failure of (1.17), Papke and Wooldridge (1996) propose a robust approach to estimate the standard errors.

Kieschnick and McCullough (2003) performed a comparison study on two datasets with common predictors, between various regression models and the quasi-likelihood approach suggested by Papke and Wooldridge (1996). The beta regression model proposed by Kieschnick and McCullough (2003) and the quasi-likelihood approach proposed by Papke and Wooldridge (1996) stood out as the best performing techniques. However, the beta regression model showed a better performance when the sample size being analyzed was small.

1.5 A Novel Generalized Estimating Equations (GEE) Approach to model Compositional Data

Since the quasi-likelihood approach relies only on the mean-variance relationship, it avoids potential problems which may arise due to distributional misspecification. If the quasi-score equations that need to be solved to obtain the required parameter estimates, are linear in the response variable being modeled, then unbiasedness and consistency of the estimator follows directly from the specification of the first moment. So a quasi-likelihood estimator is robust to the choice of a covariance structure for all the observations obtained on the response variable. Such robustness does not however extend to the estimated variance-covariance matrix of the quasi-likelihood estimator. Should the wrong covariance structure be used, the model-based standard errors obtained for the quasi-likelihood estimator will be incorrect. The quasi-likelihood estimator may, in this latter case, not retain its asymptotic efficiency (Liang and Zeger, 1986). As mentioned in the previous section, in

the case of using quasi-likelihood estimation with a logit link function and a mean-variance specification as per the Bernoulli distribution for 2-part compositional data, Papke and Wooldridge (1996) devised a robust way of estimating the standard errors. Generalized estimating equations (Liang and Zeger, 1986) is an alternative approach which may be used to analyze compositional data. GEE also relies on the specification of the first two moments. Additionally, it caters for the dependence between the response variables through the specification of a 'working' correlation matrix and estimates obtained under a GEE approach are robust to misspecification of the 'working' correlation matrix (Liang and Zeger, 1986).

The GEE approach is typically used to model longitudinal data, where the working correlation structure is used to cater for the correlation which arises between the responses that are achieved over time. Song and Tan (2000), however, use the GEE approach to estimate the parameters of a generalized linear model for longitudinal response variables with observations falling between 0 and 1. More specifically, Song and Tan (2000) assume that the marginal means depend on explanatory variables through a logit link function, and the mean-variance relationship is modeled as per the simplex distribution developed by Barndorff-Nielsen and Jørgensen (1991). Following Prentice (1988), Song and Tan (2000) introduce a second set of estimating equations to estimate the parameters making up the working correlation matrix. A drawback of using the simplex distribution (Barndorff-Nielsen and Jørgensen, 1991) to analyze compositional response variables, is that unbiasedness of the score function will fail if the assumed distribution is not the simplex. This method will thus deliver consistent estimators only when the assumed distribution holds. Furthermore, a random variable following the Barndorff-Nielsen and Jørgensen (1991) simplex distribution may only take values between 0 and 1. This method however, inspired Zhang (2013) to use Barndorff-Nielsen and Jørgensen (1991) multivariate simplex distribution to model compositional response variables, where the relationship between the mean of the response variables and the explanatory variables is modeled through a multivariate logit link. The Fisher-scoring algorithm is used to obtain maximum likelihood estimates of the model parameters. Zhang (2013) performs simulation studies by generating logistic-normally distributed and multivariate simplex distributed data. Performance of the simplex model proposed by Zhang (2013) and Aitchison (1986) approach is compared and the results obtained from the multivariate simplex model are promising. As with Aitchison (1986) regression model, however, the simplex model does not cater for any zeros in the data.

The GEE approach has also been used by Warton and Guttorp (2011) to model compositional count data, namely multivariate abundance data. Warton and Guttorp (2011) use a loglinear marginal modeling approach with mean-variance relationship specified as for an overdispersed Poisson distribution or as for a negative binomial distribution. In analogy to Aitchison (1986) approach, Warton and Guttorp (2011) choose the first component as reference component and exploit the difference in loglinear models, $\log(\mu_{ij}) - \log(\mu_{i1})$, $j \neq 1$, to model the 'compositional effects', that is, the influence of the explanatory vari-

ables on the mean response of one compositional variable in relation to the first component. An unstructured working correlation matrix is identified as the most suitable correlation matrix to use with multivariate abundance data. However, typically, the sample size of multivariate abundance data is less than the number of response variables. Warton and Guttorp (2011) expect computational issues to arise in estimating the parameters of an unstructured working correlation matrix. An independence working correlation structure is thus preferred and robust standard errors are obtained either through the use of Liang and Zeger (1986) sandwich estimator or through the use of bootstrapping.

The new approach being proposed in this thesis is directed towards modeling continuous compositional data and is also based on GEE. No detailed distributional specification will be made for the compositional response variables and the compositions will be assumed to be obtained through performing the closure operation (1.1) on a set of latent variables $\dot{Y}_j$, $j = 1, \ldots, J$, with mean-variance relationship pertaining to the family of gamma distributions with constant coefficient of variation. The choice of a notional gamma distribution as the basis of a model for compositional data bears a similarity to the method used by Gilchrist (1982) for the specific case of $J = 2$; but the approach proposed by Gilchrist (1982) is quite different, and appears to be difficult to generalize.

In a similar manner to Warton and Guttorp (2011), a loglinear model is also used in this thesis but in this case, the loglinear model is specified for the marginal mean of each latent variable. Generalized estimating equations are developed to estimate the parameters in the loglinear model and through the multiplicative nature of the model, it will be shown how this model provides an interchangeability between the latent and compositional response variables and that the differences among the parameters relating to the latent variables correspond to the compositional effects. By further considering that the fitted values corresponding to the compositional response variables should also be sum-constrained, a new system of estimating equations for compositional data is devised. This new system will be referred to by the name *hybrid* system. For the special case $J = 2$, the hybrid system is shown to be the same as Wedderburn's estimating equations (Wedderburn, 1974), which were used for the analysis of barley leaf data. A generalization of Wedderburn's system of estimating equations to $J > 2$ is then developed by constructing generalized estimating equations for a multivariate logit model and by using a working variance-covariance structure that is suitable for modeling compositional response variables. This new approach, which will be referred to by the name *generalized Wedderburn* method, is simple to implement via iterative least squares. It will also be shown how through this approach the marginal means of compositional variables may be modeled directly, and any of the problematic issues encountered by the other approaches are also avoided. In particular, the model assumptions that are used to analyze subcompositions are consistent with those used to analyze a full composition and this new approach may also be used in the presence of zeros. It might be argued that the fact that the parameter estimates obtained from analyzing a full composition using the generalized Wedderburn model are not in general the same as those obtained when a subcomposition is analyzed presents

a shortcoming of the model. We argue that, on the contrary, such a requirement is too strict for use with the generalized Wedderburn approach and might even be undesirable in general. More detail on this new approach will be given throughout the rest of the thesis.

1.6 Structure of the Thesis

This thesis is divided into six chapters.

Chapter 2 presents the theory behind the development of a multivariate logit model to be used with continuous compositional data. Estimation of the model parameters is carried out using the technique of generalized estimating equations with a working variance-covariance structure that reflects the properties of compositional variables. Different ways in which standard errors may be estimated are also explored and a new model-based variance estimator which 'borrows strength across subjects' (Liang and Zeger, 1986) is developed. Measures which are appropriate for testing the quality of fit of the multivariate logit model for compositional data are also presented.

As mentioned in Section 1.2.2, the standard methodology used to model compositional response variables is that devised by Aitchison (1982, 1986). Chapter 3 will provide more detail on Aitchison's regression method and it will show how Aitchison's regression model relates to a multiplicative regression model that is introduced in Chapter 2. Despite being two different methods, Aitchison's method and the generalized Wedderburn method have some striking similarities of form. The formal similarities of the two approaches will be presented in this chapter together with an in-depth study of the properties of estimators obtained under the two approaches. An efficiency comparison between the GEE estimator used under the generalized Wedderburn method and the ML estimator used under Aitchison's method is carried out using a small simulation study, under various sample sizes, coefficients of variation and correlation coefficients, with compositional data being generated through multivariate lognormally distributed latent variables. The generalized Wedderburn method and Aitchison's method are then compared on two widely used datasets from the compositional data literature, the Arctic Lake dataset (e.g. Aitchison, 1986; Tsagris et al., 2011; Maier, 2014) and the Foraminiferal dataset (e.g. Aitchison, 1986; Palarea-Albaladejo et al., 2007; Scealy and Welsh, 2011; Tsagris, 2015).

Chapter 4 makes some comparisons with Dirichlet models. Since the Dirichlet regression model may be specified using the same logit model that is estimated by the generalized Wedderburn approach, in Chapter 4 we first present some theoretical background on the Dirichlet regression model. A Dirichlet regression model is then fitted to the Arctic Lake dataset and the resulting fit is compared to that obtained by the generalized Wedderburn method. The estimates obtained from fitting the Dirichlet regression model to the Arctic Lake dataset are then used to generate data for a simulation study which compares the efficiency of the GEE estimator, used under the generalized Wedderburn approach, with

the ML estimator used in the Dirichlet regression model.

Chapter 5 contains a brief introduction to an early development version of the *cglm* package. This R package may be used to fit the newly proposed generalized Wedderburn method and Aitchison's multivariate regression model to compositional data. It also provides basic tools for model summary and model criticism.

Finally, Chapter 6 contains a summary of the material presented in this thesis together with some concluding comments and suggestions for further studies.

Chapter 2

A Multivariate Generalized Linear Model

2.1 Introduction

Amongst researchers of compositional data analysis, the method which is most likely to be used to model the influence of predictors on compositional response variables is that of logratio-transforming the data, assuming the distribution of the transformed data to be the multivariate normal distribution and then proceeding with using ordinary least squares estimation. However, as mentioned in Chapter 1, the logratio methodology fails when dealing with zero-valued responses. Also, the logratio methodology models the mean of the logratios, rather than the mean of the compositional response variables directly, so interpretation of regressions based on logratios is rather indirect.

In this work, a latent multiplicative regression model (MRM) is first introduced. This model is based on the consideration that in modeling compositional response variables, treating the effects and errors as multiplicative on the untransformed components is more suitable than treating them as additive. Also, rather than modeling transformed data, the MRM transforms the model expectations, in the already-familiar way that generalized linear models represent an alternative to data transformation prior to linear modelling.

The fact that a multiplicative model is used to model compositional data is based on the analogy of the operation of perturbation [1] (Aitchison, 1986), which is a multiplicative operation in the simplex, with the operation of translation [2], the latter being an additive

[1] **Definition 2.1.1.** *The perturbation between any two J-part compositions $\boldsymbol{Y}^*$ and $\boldsymbol{Y}$ is defined by*

$$\boldsymbol{Y}^* \oplus \boldsymbol{Y} = C(Y_1^* Y_1, \ldots, Y_J^* Y_J)$$

where $\oplus$ is the notation that is typically used to denote the perturbation operation and $C(\cdot)$ denotes the closure operation that has been defined in (1.1).

[2] Consider two compositions $\mathbf{Y}$ and $\mathbf{Y}^*$ which are related by $\mathbf{Y} = \mathbf{p} \oplus \mathbf{Y}^*$. Let $\mathbf{W}$, $\mathbf{P}^*$ and $\mathbf{W}^*$

operation in the real space.

The motivation for the MRM modeling the mean on the original scale comes from Firth (1987, 1988), who has shown that modeling the mean on the original scale through a multiplicative model rather than on the log-transformed data might yield better efficiency of the estimators, as well as overcoming the aforementioned problems of the analysis of logarithms.

The latent multiplicative regression model (MRM) is presented in Section 2.2. A brief note on identifying the parameters of the MRM is presented in Section 2.3. Section 2.4 focuses on parameter estimation. Since only the first two moments of the latent variables underlying the compositional response variables and no further distributional assumption is made in the specification of the latent MRM, it will be shown how quasi-likelihood estimation may be used to estimate the parameters in the MRM. Details on the general technique of quasi-likelihood estimation and properties of the quasi-likelihood estimator are provided in Section 2.4.1. Section 2.4.2 then explains how quasi-likelihood methods may be applied to estimate the parameters in the MRM. A quasi-likelihood estimator is robust to the specification of a covariance structure but this robustness does not extend to the estimated variance-covariance matrix of the quasi-likelihood estimator. This drawback is overcome through using the technique of generalized estimating equations (GEE). The technique of generalized estimating equations uses the mean-variance specification of quasi-likelihood estimation but it is also able to cater for any correlation that may arise between the observed variables by introducing a working correlation matrix. Section 2.4.3 shows how generalized estimating equations may be applied to estimate the parameters in the multiplicative regression model. It will also be shown that the generalized least squares estimator which is used to estimate the parameters of interest is invariant under different dispersion and correlation parameters. Independence estimating equations with equal dispersion parameters may thus be used to estimate the model parameters, which makes this system of estimating the parameters very appealing. The problem with using such a system, however, is that the sum of the estimated means is not constrained to be equal to 1. Compositional response variables are sum constrained, so their estimated means should be constrained accordingly. An alternative new system of estimating equations, referred to by the name *hybrid* is developed in Section 2.4.5. The hybrid system retains the invariance property of the generalized least squares estimator for the parameters of interest whilst

---

denote $J-1$-vectors whose components are the logratios defined as $\log\left(\frac{Y_j}{Y_J}\right)$, $\log\left(\frac{p_j}{p_J}\right)$ and $\log\left(\frac{Y_j^*}{Y_J^*}\right)$, for $j = 1, \ldots, J-1$, respectively. Then

$$\begin{aligned}
\mathbf{P}^* + \mathbf{W}^* &= \left(\left[\log\left(\frac{Y_1 Y_1^{*-1}}{Y_J Y_J^{*-1}}\right) + \log\left(\frac{Y_1^*}{Y_J^*}\right)\right], \ldots, \left[\log\left(\frac{Y_{J-1} Y_{J-1}^{*-1}}{Y_J Y_J^{*-1}}\right) + \log\left(\frac{Y_{J-1}^*}{Y_J^*}\right)\right]\right) \\
&= \ldots \\
&= \left(\log\left(\frac{Y_1}{Y_J}\right), \ldots, \log\left(\frac{Y_{J-1}}{Y_J}\right)\right) \\
&= \mathbf{W},
\end{aligned}$$

which is in the form of a translation in $\mathbb{R}^{J-1}$.

also imposing the sum constraint on the estimated means. In Section 2.5, the estimating equations obtained under the hybrid system when $J = 2$ are also shown to be the same as Wedderburn's estimating equations (Wedderburn, 1974), which were used for the analysis of barley leaf data. Through this equivalence, the hybrid system for $J = 2$ inherits all the desirable properties of Wedderburn's quasi-likelihood estimator. Based on this development, in Section 2.6 a generalization of Wedderburn's system of estimating equations to $J > 2$ is sought and developed by constructing generalized estimating equations for a multivariate logit model. A working variance-covariance structure that is suitable for modeling compositional response variables is then identified in Section 2.7. Different ways in which standard errors may be estimated are explored in Section 2.8. A new estimator of the standard errors which 'borrows strength across subjects' (Liang and Zeger, 1986) is developed in Section 2.8.2. The final section first presents different measures that are used in testing quality of fit in a typical GEE analysis. Subsequently, measures which are appropriate for testing the quality of fit of our logit model for compositional data are presented.

2.2 The Latent Multiplicative Regression Model (MRM)

For a latent MRM, suppose that for a sample of size n and a set of predictors $X_1, \ldots, X_p$, the random variables $\dot{Y}_{ij}$, $i = 1, \ldots, n$, $j = 1, \ldots, J$ are latent variables that are modeled multiplicatively as

$$\dot{Y}_{ij} = m_i \left(\dot{\theta}_i, \boldsymbol{\beta}_j \right) E_{ij} \tag{2.1}$$

where the function m_i for the i^{th} case is defined as

$$m_i \left(\theta, \boldsymbol{\beta} \right) = \exp \left(\theta + \mathbf{x}_i^{'} \boldsymbol{\beta} \right). \tag{2.2}$$

Given values of $\dot{\theta}_i$, $\boldsymbol{\beta}_j$ and $\mathbf{x}_i$, $m_i \left(\dot{\theta}_i, \boldsymbol{\beta}_j \right)$ is the conditional expectation of $\dot{Y}_{ij}$ which expresses dependence on predictor variables X_k, $k = 1, \ldots, p$, through the log link, E_{ij} is the error term associated with $\dot{Y}_{ij}$, $\dot{\theta}_i$ is an unknown (nuisance) parameter that will need to be estimated, $\boldsymbol{\beta}_j$ is a $(p+1)$-vector of coefficients that also needs to be estimated, $\mathbf{x}_i$ is a $(p+1)$-vector with $x_{i0} = 1$ since the first element corresponds to the intercept and the remaining elements correspond to the observations obtained by the ith case in the sample on $X_1, \ldots, X_p$. Note that the intercept is introduced in the model since we are not assuming that the explanatory variables $X_1, \ldots, X_p$ can take some special zero value.

The error vectors $\mathbf{E}_i = (E_{i1}, \ldots, E_{iJ})^{'}$ are assumed to be independent of one another, with

$$E \left(\mathbf{E}_i \right) = \mathbf{1}. \tag{2.3}$$

Since the variance-covariance matrix of $\mathbf{E}_i$ is the same for all i it will be denoted by $\dot{\boldsymbol{\Sigma}}$

and will be taken to be of the form

$$\dot{\boldsymbol{\Sigma}} = \phi \boldsymbol{\Omega}^{\frac{1}{2}} \mathbb{W} \boldsymbol{\Omega}^{\frac{1}{2}}, \tag{2.4}$$

where ϕ is a common dispersion parameter, $\boldsymbol{\Omega} = \text{diag}\,(\omega_1, \ldots, \omega_J)$ with $\omega_1, \ldots, \omega_J$ being relative dispersion parameters attributed to the J random variables $E_{i1}, \ldots, E_{iJ}$ and $\mathbb{W}$ is the correlation matrix.

Letting $\alpha_{jj'}$ denote the elements in the $J \times J$ matrix $\mathbb{W}$ and using equations (2.1) and (2.4), it follows that

$$\begin{aligned} E\left(\dot{Y}_{ij}\right) = m_i\left(\dot{\theta}_i, \boldsymbol{\beta}_j\right), \qquad \text{Var}\left(\dot{Y}_{ij}\right) = \phi\omega_j\left[m_i\left(\dot{\theta}_i, \boldsymbol{\beta}_j\right)\right]^2 \\ \text{Cov}\left(\dot{Y}_{ij}, \dot{Y}_{ij'}\right) = \phi\sqrt{\omega_j}\sqrt{\omega_{j'}} m_i\left(\dot{\theta}_i, \boldsymbol{\beta}_j\right) m_i\left(\dot{\theta}_i, \boldsymbol{\beta}_{j'}\right) \alpha_{jj'}. \end{aligned} \tag{2.5}$$

It may be noticed that the MRM is specified in terms of the variables $\dot{Y}_{i1}, \ldots, \dot{Y}_{iJ}$. At this stage this might seem confusing, since our aim is that of modeling compositional variables $Y_{i1}, \ldots, Y_{iJ}$. However, recall that compositional variables are obtained as a result of perfoming the closure operation (1.1) on $\dot{Y}_{i1}, \ldots, \dot{Y}_{iJ}$. The variables $\dot{Y}_{i1}, \ldots, \dot{Y}_{iJ}$ will thus be considered as latent variables through which we obtain the observed compositional variables $Y_{i1}, \ldots, Y_{iJ}$. To get a better insight into why the model for the latent variables may be used to model the compositional response variables, consider the following.

Let the latent variable $\dot{Y}_{ij}$ and its compositional counterpart Y_{ij} be related through the equation

$$\dot{Y}_{ij} = c_i Y_{ij}. \tag{2.6}$$

On using the closure operation (1.1), it may be noted that c_i is some unknown positive constant defined by

$$c_i = \dot{Y}_{i1} + \ldots + \dot{Y}_{iJ}. \tag{2.7}$$

Since the value of c_i is related solely to case i, if its value were to change to say c_i^*, the only change in the MRM (2.1) would be in the values of the parameters $\left(\dot{\theta}_1, \ldots, \dot{\theta}_n\right)$, which are nuisance parameters. They have been introduced in model (2.1) to cater for the rescaling of $\dot{Y}_{1j}, \ldots, \dot{Y}_{nj}$. So the fact that a change in c_i leads to a change in the values of $\left(\dot{\theta}_1, \ldots, \dot{\theta}_n\right)$ should not be considered a problem. The parameters of interest are $(\boldsymbol{\beta}_1, \ldots, \boldsymbol{\beta}_J)$ and by changing the value of c_i, the values of $(\boldsymbol{\beta}_1, \ldots, \boldsymbol{\beta}_J)$ are not affected. An advantage of the just mentioned, is that the value of the constants $c_1, \ldots, c_n$ may be taken to be any positive value of choice, including the value of unity. On taking all c_is to be equal to one, the latent variables $\dot{Y}_{ij}$ will be equal to the compositional variables Y_{ij}, so in practice, the MRM (2.1) may be used to model the compositional variables Y_{ij} directly.

The idea behind using an interchangeability between the compositional (constrained) Y_{ij}

and the unconstrained $\dot{Y}_{ij}$ may be viewed as analogous to the 'Poisson trick' (Palmgren, 1981; Kosmidis and Firth, 2011) which gives the interchangeability of the Poisson distribution and the multinomial distribution for loglinear models and multinomial logit models respectively, since the multinomial distribution is invoked by conditioning on the observed marginal totals for the predictors of a Poisson sampling model for a contingency table. In the approach proposed here for compositional data, the compositional response variables arise out of the closure operation on the latent variables and despite the fact that no detailed distributional specifications will be made in this novel approach, the analogy with a multinomial logit model may be appreciated through considering equation (2.8) which follows shortly.

The interchangeability between Y_{ij}s and $\dot{Y}_{ij}$s is also obtained in the estimation procedure. In order to estimate the model parameters in (2.1), focus will be directed towards modeling the mean of $\dot{Y}_{ij}$ through the latent multiplicative regression model defined in equations (2.1) and (2.5). More details on this will be given in Section 2.4.3. Some issues related with identification of the model parameters need to be discussed before delving into how to estimate the model parameters.

2.3 Identification of the Parameters of the Latent MRM

Due to the sum constraint, when dealing with compositional data, it is the components making up a composition taken in relation to some reference component (the ratios introduced by Aitchison (1986)), that are of importance, rather than the components themselves. Without loss of generality, taking the last component as the reference component, the logratios of expectations, as defined through the MRM, take the form

$$\log\left[E\left(\dot{Y}_{ij}\right)\right] - \log\left[E\left(\dot{Y}_{iJ}\right)\right] = \left(\beta_{j0} - \beta_{J0}\right) x_{i0} + \left(\beta_{j1} - \beta_{J1}\right) x_{i1} + \cdots + \left(\beta_{jp} - \beta_{Jp}\right) x_{ip}. \tag{2.8}$$

As a consequence of (2.8), without imposing any constraints on the parameters of the latent MRM, parameter identification is only possible for differences $(\beta_{jk} - \beta_{Jk})$. If any constant is added to $\beta_{j0}, \ldots, \beta_{jp}$ and/or $\beta_{J0}, \ldots, \beta_{Jp}$ in the MRM (2.1), the distribution of the composition will not be changed. So to achieve a 1-1 mapping between parameter values and distributions, a reparametrization on the β parameters is needed. In Section 2.5, it will be shown that through the use of a newly developed system of estimating equations, compositional data may actually be modeled through a logit model with the model coefficients being the differences $(\beta_{jk} - \beta_{Jk})$ and focus will be directed towards the differences $(\beta_{jk} - \beta_{Jk})$. The theoretical background that needs to be prepared prior to that section will however be based on the full set of β parameters.

Also note that without imposing a constraint on the dispersion parameters $\omega_1, \ldots, \omega_J$ or ϕ, it will not be possible to identify all dispersion parameters. Without loss of generality, in this thesis, $\omega_1, \ldots, \omega_J$ will be assumed to be *relative* dispersion parameters which sum

to J.

2.4 Estimating the Parameters of the Latent MRM

2.4.1 Quasi-Likelihood Estimation

Estimation of the parameters for the proposed latent MRM (2.1) may be carried out using an adaptation of quasi-likelihood estimation. Prior to explaining how quasi-likelihood estimation may be slightly modified to be used to estimate the parameters of the multiplicative regression model, an overview of the quasi-likelihood estimation technique (Wedderburn, 1974; McCullagh, 1983; Firth, 1993b) is first presented.

Suppose that for a sample of size n, an n-vector of response variables $\dot{\mathbf{Y}}$ satisfies

$$E\left(\dot{\mathbf{Y}}\right) = \mathbf{m}\left(\boldsymbol{\beta}\right) \quad \text{and} \quad \text{Var}\left(\dot{\mathbf{Y}}\right) = \phi\dot{\mathbb{V}}\left(\mathbf{m}\left(\boldsymbol{\beta}\right)\right), \tag{2.9}$$

where the random variables in $\dot{\mathbf{Y}}$ are assumed to be independent of each other, $\boldsymbol{\beta}$ is a vector of unknown regression parameters that will need to be estimated and $\mathbf{m}\left(\boldsymbol{\beta}\right)$ has components $m_i\left(\boldsymbol{\beta}\right)$, $i = 1, \ldots, n$, where the functions $m_i\left(\boldsymbol{\beta}\right)$ express dependence on predictor variables X_k, $k = 1, \ldots, p$, through a generalized linear model $g\left(m_i\left(\boldsymbol{\beta}\right)\right) = \mathbf{x}_i'\boldsymbol{\beta}$, where $g\left(\cdot\right)$ is a specified link function. Thus, the function $m_i\left(\boldsymbol{\beta}\right)$ is defined by

$$m_i\left(\boldsymbol{\beta}\right) = g^{-1}\left(\mathbf{x}_i^{'}\boldsymbol{\beta}\right). \tag{2.10}$$

Also, in (2.9), ϕ is the dispersion parameter and $\dot{\mathbb{V}}\left(\mathbf{m}\left(\boldsymbol{\beta}\right)\right)$ is a symmetric, positive-definite matrix of known functions of unknown means $m_i\left(\boldsymbol{\beta}\right)$ of $\dot{Y}_i$. Functions of the mean that are used to make up $\dot{\mathbb{V}}\left(\mathbf{m}\left(\boldsymbol{\beta}\right)\right)$ may take a form which does not necessarily correspond to that of a specific distribution. As an example of this consider the function $\left(m_i\left(\boldsymbol{\beta}\right)\right)^2\left(1 - m_i\left(\boldsymbol{\beta}\right)\right)^2$, which is the variance function that is used by Wedderburn (1974) in his study on the proportion of barley leaf area that was infected with *Rhynchosporium secalis*.

Let $\dot{\mathbb{D}}$ denote an $n \times (p+1)$ matrix of derivatives with components

$$\frac{\partial m_i}{\partial \beta_k} = \frac{\partial m_i}{\partial \eta_i}\frac{\partial \eta_i}{\partial \beta_k} = \frac{\partial m_i}{\partial \eta_i}x_{ik},$$

where m_i stands for $m_i\left(\boldsymbol{\beta}\right)$, η_i stands for $\eta_i\left(\boldsymbol{\beta}\right)$ and $\eta_i\left(\boldsymbol{\beta}\right) = \mathbf{x}_i^{'}\boldsymbol{\beta}$. Estimation of the model parameters may then proceed by setting up the quasi-score vector

$$\mathbf{U} = \dot{\mathbb{D}}'\left[\text{Var}\left(\dot{\mathbf{Y}}\right)\right]^{-1}\left(\dot{\mathbf{Y}} - \mathbf{m}\left(\boldsymbol{\beta}\right)\right), \tag{2.11}$$

which yields the quasi-likelihood equations

$$\mathbf{U} = \dot{\mathbb{D}}' \left[\mathrm{Var}\left(\dot{\mathbf{Y}} \right) \right]^{-1} \left(\dot{\mathbf{Y}} - \mathbf{m}\left(\boldsymbol{\beta} \right) \right) = \mathbf{0}. \tag{2.12}$$

An estimate of $\boldsymbol{\beta}$ is obtained through solving equations (2.12). Since ϕ may easily be taken out of the equation (2.12), the estimates of $\boldsymbol{\beta}$ do not depend on the value of the dispersion parameter. In fact, ϕ may be estimated using a moment estimator based on Pearson residuals, at the very end, that is, once convergence has been reached in the estimates of $\boldsymbol{\beta}$.

So for quasi-likelihood estimation to be carried out, the only specifications that are required are those of the first and second moment of the response vector. In this respect, the appropriateness of the term 'likelihood' in this estimation technique's name might be questioned. However, Wedderburn (1974) used the term quasi-likelihood because of the similarities between the log quasi-likelihood function and the log-likelihood function or more specifically, in the behaviour of the vector $\mathbf{U}$ and that of the likelihood score vector $\frac{\partial l}{\partial \boldsymbol{\beta}}$. Effectively, under (2.9), the properties

$$E\left(\mathbf{U} \right) = \mathbf{0} \quad \text{and} \quad \mathrm{Var}\left(\mathbf{U} \right) = -E\left(\frac{\partial \mathbf{U}}{\partial \boldsymbol{\beta}} \right) = \dot{\mathbb{D}}' \left[\mathrm{Var}\left(\dot{\mathbf{Y}} \right) \right]^{-1} \dot{\mathbb{D}} \tag{2.13}$$

hold if $\mathbf{U}$ is a quasi-score function and also if it is the score vector obtained from a regular log-likelihood function.

Additionally, asymptotic properties of consistency, normality and unbiasedness of maximum likelihood estimators also hold for quasi-likelihood estimators (McCullagh, 1983). This latter property is due to the fact that first-order asymptotic theory of maximum likelihood estimators and related inference procedures are based on properties (2.13). Consequently, under the usual limiting conditions on the eigenvalues of $\mathrm{Var}\left(\mathbf{U} \right)$ [3] and other theoretical conditions[4], the inverse of the variance-covariance matrix $\mathrm{Var}\left(\mathbf{U} \right)$ plays the same role as the Fisher information matrix for regular likelihood functions. Thus

$$\mathrm{Var}\left(\widehat{\boldsymbol{\beta}} \right) \approx \left[\mathrm{Var}\left(\mathbf{U} \right) \right]^{-1} = \left(\dot{\mathbb{D}}' \left[\mathrm{Var}\left(\dot{\mathbf{Y}} \right) \right]^{-1} \dot{\mathbb{D}} \right)^{-1}. \tag{2.14}$$

Moreover, under the mean-variance specification (2.9), quasi-likelihood estimators have been shown to maximize asymptotic efficiency amongst unbiased estimating equations that are linear in $\dot{\mathbf{Y}}$ (Firth, 1987, 1993b). Firth (1987) has also shown that quasi-likelihood estimators preserve 'fairly high efficiency under moderate departures' from linear exponential family distributions.

[3] 'Eigenvalues of $\mathrm{Var}\left(\mathbf{U} \right)$ should tend to infinity for all $\boldsymbol{\beta}$ in an open neighbourhood of the true parameter point' (McCullagh and Nelder, 1989, p. 333).

[4] 'Roughly speaking, as the sample size $n \to \infty$, the quasi-score vector $\mathbf{U}$ should be asymptotically normal' (McCullagh and Nelder, 1989, p. 333).

2.4.2 Applying Quasi-Likelihood Estimation to the MRM

In this section, it will be shown how quasi-likelihood estimation may be used to estimate the parameters of the MRM. To be able to follow how the material presented in Section 2.4.1 will be used in this section, it should be noted that since compositional data takes the form of an $n \times J$ matrix of observations, the n-vector of response variables $\dot{\mathbf{Y}}$ that has been used in the previous section, will now be taken to be an nJ-vector of response variables. More specifically, $\dot{\mathbf{Y}} = \left(\dot{\mathbf{Y}}_1^{'}, \ldots, \dot{\mathbf{Y}}_n^{'}\right)^{'}$ where for $i = 1, \ldots, n$, $\dot{\mathbf{Y}}_i = \left(\dot{Y}_{i1}, \ldots, \dot{Y}_{iJ}\right)^{'}$. Also, the vector of parameters that needs to be estimated will now be denoted by $\boldsymbol{\beta}^{+}$ where $\boldsymbol{\beta}^{+} = \left(\dot{\boldsymbol{\theta}}^{'}, \boldsymbol{\beta}^{'}\right)^{'}$; here $\dot{\boldsymbol{\theta}} = \left(\dot{\theta}_1, \ldots, \dot{\theta}_n\right)^{'}$ and $\boldsymbol{\beta}$ is a $J\,(p+1)$-vector given by $\boldsymbol{\beta} = \left(\boldsymbol{\beta}_1^{'}, \ldots, \boldsymbol{\beta}_J^{'}\right)^{'}$, where for $(j = 1, \ldots, J)$, $\boldsymbol{\beta}_j = (\beta_{j0}, \ldots, \beta_{jp})^{'}$. Then, using (2.5), it may be seen that

$$E\left(\dot{\mathbf{Y}}\right) = \mathbf{m}\left(\boldsymbol{\beta}^{+}\right) \quad \text{and} \quad \text{Var}\left(\dot{\mathbf{Y}}\right) = \phi\boldsymbol{\Omega}\dot{\mathbb{V}}\left(\mathbf{m}\left(\boldsymbol{\beta}^{+}\right)\right), \tag{2.15}$$

where the random variables in $\dot{\mathbf{Y}}$ are assumed to be independent of each other, $\mathbf{m}\left(\boldsymbol{\beta}^{+}\right)$ has components $m_i\left(\dot{\theta}_i, \boldsymbol{\beta}_j\right)$ where $m_i\left(\theta, \boldsymbol{\beta}\right)$ is the function defined in (2.2), $\dot{\mathbb{V}}\left(\mathbf{m}\left(\boldsymbol{\beta}^{+}\right)\right)$ is an $nJ \times nJ$ variance-covariance matrix and the variances making up $\dot{\mathbb{V}}\left(\mathbf{m}\left(\boldsymbol{\beta}^{+}\right)\right)$ are $\left[m_i\left(\dot{\theta}_i, \boldsymbol{\beta}_j\right)\right]^2$, so $\dot{\mathbb{V}}\left(\mathbf{m}\left(\boldsymbol{\beta}^{+}\right)\right)$ is a symmetric, positive-definite matrix of known functions of unknown means $m_i\left(\dot{\theta}_i, \boldsymbol{\beta}_j\right)$ of $\dot{Y}_{ij}$.

The relationship between the first two moments for the MRM, presented in (2.15), is in the same form as that presented in (2.9). One main difference between (2.9) and (2.15) is that for the latent MRM, the dispersion parameter is allowed to take different values across the J components. So rather than having one dispersion parameter ϕ, for the latent MRM this is generalized to the matrix $\phi\boldsymbol{\Omega}$, $\boldsymbol{\Omega}$ being a diagonal matrix whose elements are $\omega_1, \ldots, \omega_J$, that also need to be estimated.

Under the latent MRM, let $\dot{\mathbb{D}} = \left(\dot{\mathbb{D}}_1^{'}, \ldots, \dot{\mathbb{D}}_n^{'}\right)^{'}$ where $\dot{\mathbb{D}}_i$ is a $J \times (J\,(p+1) + n)$ matrix whose elements are

$$\frac{\partial m_i\left(\dot{\theta}_i, \boldsymbol{\beta}_j\right)}{\partial \beta_{jk}} = m_i\left(\dot{\theta}_i, \boldsymbol{\beta}_j\right) x_{ik} \quad \text{and} \quad \frac{\partial m_i\left(\dot{\theta}_i, \boldsymbol{\beta}_j\right)}{\partial \dot{\theta}_i} = m_i\left(\dot{\theta}_i, \boldsymbol{\beta}_j\right). \tag{2.16}$$

The quasi-score vector takes the form $\mathbf{U} = (\mathbf{U}_1, \ldots, \mathbf{U}_{J+1})^{'}$, where

- $\mathbf{U}_j$, is a $p+1$-vector that involves the derivatives of $m_i\left(\dot{\theta}_i, \boldsymbol{\beta}_j\right)$ with respect to $(\beta_{j0}, \ldots, \beta_{jp})$, for $j = 1, \ldots, J$,
- $\mathbf{U}_{J+1}$ is an n-vector that involves the derivatives of $m_i\left(\dot{\theta}_i, \boldsymbol{\beta}_j\right)$ with respect to $\left(\dot{\theta}_1, \ldots, \dot{\theta}_n\right)$.

The $(J(p+1)+n)$ quasi-score equations based on the latent responses $\dot{Y}_{ij}$ are then

$$U_{js} = \frac{1}{\phi\omega_j}\sum_{i=1}^{n}\left(\frac{\dot{Y}_{ij}}{m_i\left(\dot{\theta}_i, \boldsymbol{\beta}_j\right)} - 1\right)x_{i,s-1} = 0 \quad \text{for } j = 1,\ldots,J, s = 1,\ldots,p+1 \tag{2.17}$$

$$U_{J+1,i} = \frac{1}{\phi}\sum_{j=1}^{J}\left[\frac{1}{\omega_j}\left(\frac{\dot{Y}_{ij}}{m_i\left(\dot{\theta}_i, \boldsymbol{\beta}_j\right)} - 1\right)\right] = 0 \quad \text{for } i = 1,\ldots,n. \tag{2.18}$$

Now, as mentioned already, one of the advantages of using quasi-likelihood estimation to estimate the parameters of the MRM is that it requires only the mean-variance relationship to be stipulated, so it avoids potential problems in the estimators resulting due to distributional misspecification. A crucial aspect of using such a technique to estimate the parameters of the latent MRM, as will be shown shortly, is also that the latent variables $\dot{Y}_{ij}$ in the quasi-score equations may be replaced with the compositional response variables Y_{ij} without affecting the resulting estimates of $\beta_{10},\ldots,\beta_{Jp}$.

Using (2.6), the quasi-score equation (2.17) may be written in terms of Y_{ij}s as follows:

$$U_{js} = \frac{1}{\phi\omega_j}\sum_{i=1}^{n}\left(\frac{c_i Y_{ij}}{m_i\left(\dot{\theta}_i, \boldsymbol{\beta}_j\right)} - 1\right)x_{i,s-1} = 0, \tag{2.19}$$

where

$$\begin{aligned} c_i &= \dot{Y}_{i1} + \cdots + \dot{Y}_{iJ} \\ &= \exp\left(\dot{\theta}_i\right)\left[\exp\left(\boldsymbol{\beta}_1'\mathbf{x}_i\right)E_{i1} + \cdots + \exp\left(\boldsymbol{\beta}_J'\mathbf{x}_i\right)E_{iJ}\right]. \end{aligned} \tag{2.20}$$

Now, on changing scale, from c_i to say c_i^*,

$$c_i^* = \exp\left(\dot{\theta}_i^*\right)\left[\exp\left(\boldsymbol{\beta}_1'\mathbf{x}_i\right)E_{i1} + \cdots + \exp\left(\boldsymbol{\beta}_J'\mathbf{x}_i\right)E_{iJ}\right]. \tag{2.21}$$

Also, on changing the scale, the value of the mean $m_i\left(\dot{\theta}_i, \boldsymbol{\beta}_j\right) = \exp\left(\dot{\theta}_i + \mathbf{x}_i'\boldsymbol{\beta}_j\right)$ will also change due to a change in the value of $\dot{\theta}_i$. Consequently, the value of

$$\frac{c_i Y_{ij}}{m_i\left(\dot{\theta}_i, \boldsymbol{\beta}_j\right)} = \frac{\exp\left(\dot{\theta}_i\right)\left[\exp\left(\boldsymbol{\beta}_1'\mathbf{x}_i\right)E_{i1} + \cdots + \exp\left(\boldsymbol{\beta}_J'\mathbf{x}_i\right)E_{iJ}\right]Y_{ij}}{\exp\left(\dot{\theta}_i\right)\exp\left(\mathbf{x}_i'\boldsymbol{\beta}_j\right)} \tag{2.22}$$

is equal to

$$\frac{c_i^* Y_{ij}}{m_i\left(\dot{\theta}_i, \boldsymbol{\beta}_j\right)} = \frac{\exp\left(\dot{\theta}_i^*\right)\left[\exp\left(\boldsymbol{\beta}_1' \mathbf{x}_i\right) E_{i1} + \cdots + \exp\left(\boldsymbol{\beta}_{J-1}' \mathbf{x}_i\right) E_{i,J-1} + E_{iJ}\right] Y_{ij}}{\exp\left(\dot{\theta}_i^*\right) \exp\left(\mathbf{x}_i' \boldsymbol{\beta}_j\right)}, \quad (2.23)$$

because of the cancellation of the terms $\exp(\theta_i)$ and $\exp(\theta_i^*)$ in equations (2.22) and (2.23) respectively. In (2.22) and (2.23), the change that occurs in $\dot{\theta}_i$ in the numerator cancels with the change that also occurs in the denominator.

So whichever value of c_i is used, $\sum_{i=1}^{n}\left(\frac{c_i Y_{ij}}{m_i\left(\dot{\theta}_i, \boldsymbol{\beta}_j\right)} - 1\right) x_{i,s-1}$ will always have the same value since it will always be free of $\dot{\theta}_i$ and will lead to the same estimates of $\beta_{10}, \ldots, \beta_{Jp}$. It therefore follows that the value of c_i may actually be set equal to 1 and in so doing the quasi-score equations will be in terms of the observed compositional response variables Y_{ij} rather than in terms of the unobserved $\dot{Y}_{ij}$.

The quasi-score equations may also still be used if any zeros are present in the data. Also, the quasi-score equations that result for the latent MRM are linear in $\dot{Y}_{ij}$ so consistency of the estimator follow directly from the specification of the first moment. So a quasi-likelihood estimator is robust to the choice of a covariance structure $\dot{\mathbb{V}}\left(\mathbf{m}\left(\boldsymbol{\beta}^+\right)\right)$ under both independence and non-independence of $\dot{Y}_{i1}, \ldots, \dot{Y}_{iJ}$. The problem with using this estimation technique is however that robustness does not extend to the estimated variance-covariance matrix of the quasi-likelihood estimator. Should the wrong covariance structure be used, the model-based standard errors obtained for the quasi-likelihood estimator will be incorrect. The quasi-likelihood estimator may in general, in this latter case, not retain its asymptotic efficiency (Liang and Zeger, 1986).

So, if there is reason to believe that random variables $\dot{Y}_{i1}, \ldots, \dot{Y}_{iJ}$ are non-independent, a covariance structure which caters for this dependence should be used. In dealing with latent variables $\dot{Y}_{ij}$ from which compositional variables Y_{ij} arise, it is more realistic to consider the $\dot{Y}_{ij}$s to be related. The dependence between these latent variables may be catered for by using the technique of generalized estimating equations (GEE) which necessitates the specification of a 'working' correlation matrix.

A description of how the method of generalized estimating equations (Liang and Zeger, 1986) is applied for the estimation of the parameters in the latent multiplicative model follows. All of the advantages of using quasi-likelihood estimation will continue to hold under the GEE framework, even if the working correlation matrix is misspecified. Let $\widehat{\text{Var}\left(\widehat{\boldsymbol{\beta}}^+\right)}$ denote the estimator of the variance-covariance matrix $\text{Var}\left(\widehat{\boldsymbol{\beta}}^+\right)$ achieved under a typical GEE procedure (more detail on how to estimate standard errors will be given in Section 2.8). The consistency of $\widehat{\text{Var}\left(\widehat{\boldsymbol{\beta}}^+\right)}$ is determined by the specification of the mean model, not by the specification of the working correlation matrix. This makes the GEE estimator more robust than the quasi-likelihood estimator.

2.4.3 Applying Generalized Estimating Equations to the Latent MRM

Consider once again the proposed latent multiplicative regression model (2.5) and consider the nJ-vector of latent variables $\dot{\mathbf{Y}}$ where $\dot{\mathbf{Y}} = \left(\dot{\mathbf{Y}}_1^{'}, \ldots, \dot{\mathbf{Y}}_n^{'}\right)^{'}$. For $i = 1, \ldots, n$, under the generalized estimating equations (GEE) framework proposed by Liang and Zeger (1986), the variance-covariance matrix of $\dot{\mathbf{Y}}_i$ is assumed to take the form

$$\text{Var}\left(\dot{\mathbf{Y}}_i\right) = \phi \mathbb{A}_i^{\frac{1}{2}} \mathbb{W}\left(\boldsymbol{\alpha}\right) \mathbb{A}_i^{\frac{1}{2}} \tag{2.24}$$

where ϕ is the scalar dispersion parameter, $\mathbb{A}_i$ is a $J \times J$ diagonal matrix whose elements are $\frac{\text{Var}\left(\dot{Y}_{ij}\right)}{\phi}$, $\text{Var}\left(\dot{Y}_{ij}\right)$ being a function of the mean of $\dot{Y}_{ij}$, and $\mathbb{W}\left(\boldsymbol{\alpha}\right)$ is a $J \times J$ 'working' correlation matrix that is fully specified by the vector of parameters $\boldsymbol{\alpha}$.

Based on Paik (1992), the variance-covariance structure (2.24) will be slightly modified for the latent MRM to allow unequal dispersion parameters $(\phi\omega_1, \ldots, \phi\omega_J)$ for the J latent variables $\dot{Y}_1, \ldots, \dot{Y}_J$. The variance-covariance structure that will be considered for the latent MRM is thus

$$\text{Var}\left(\dot{\mathbf{Y}}_i\right) = \phi \boldsymbol{\Omega}^{\frac{1}{2}} \mathbb{A}_i^{\frac{1}{2}} \mathbb{W}\left(\boldsymbol{\alpha}\right) \mathbb{A}_i^{\frac{1}{2}} \boldsymbol{\Omega}^{\frac{1}{2}} \tag{2.25}$$

where $\boldsymbol{\Omega} = \text{diag}\left(\omega_1, \ldots, \omega_J\right)$, $\mathbb{A}_i$ is a $J \times J$ diagonal matrix defined by

$$\mathbb{A}_i = \text{diag}\left(\left[m_i\left(\dot{\theta}_i, \boldsymbol{\beta}_1\right)\right]^2, \ldots, \left[m_i\left(\dot{\theta}_i, \boldsymbol{\beta}_J\right)\right]^2\right) \tag{2.26}$$

and $\mathbb{W}\left(\boldsymbol{\alpha}\right)$ is a $J \times J$ working correlation matrix where $\alpha_{jj} = 1$.

Let $\dot{\mathbb{D}} = \left(\dot{\mathbb{D}}_1^{'}, \ldots, \dot{\mathbb{D}}_n^{'}\right)^{'}$ where $\dot{\mathbb{D}}_i$ is a $J \times \left(J\left(p+1\right) + n\right)$ matrix whose elements are

$$\frac{\partial m_i\left(\dot{\theta}_i, \boldsymbol{\beta}_j\right)}{\partial \beta_{jk}} = m_i\left(\dot{\theta}_i, \boldsymbol{\beta}_j\right) x_{ik} \quad \text{and} \quad \frac{\partial m_i\left(\dot{\theta}_i, \boldsymbol{\beta}_j\right)}{\partial \dot{\theta}_i} = m_i\left(\dot{\theta}_i, \boldsymbol{\beta}_j\right), \tag{2.27}$$

and let $\text{Var}\left(\dot{\mathbf{Y}}\right)$ be an nJ by nJ block diagonal matrix with variance-covariance matrices $\text{Var}\left(\dot{\mathbf{Y}}_i\right)$ on the diagonal. Then, as per Liang and Zeger (1986), estimation of the latent MRM parameters using GEE proceeds by setting up the estimating equations

$$\dot{\mathbb{D}}' \left[\text{Var}\left(\dot{\mathbf{Y}}\right)\right]^{-1} \left(\dot{\mathbf{Y}} - \mathbf{m}\left(\boldsymbol{\beta}^{+}\right)\right) = \mathbf{0}, \tag{2.28}$$

where if the correlation matrix $\mathbb{W}\left(\boldsymbol{\alpha}\right)$ is set equal to the identity matrix, (2.28) reduce to the quasi-likelihood equations (2.12), showing that quasi-likelihood equations may be viewed as a special case of generalized estimating equations. The main difference between the two sets of equations lies in the specification of $\text{Var}\left(\dot{\mathbf{Y}}\right)$. The quasi-likelihood

equations are solved to get an estimate of $\boldsymbol{\beta}^+$, ϕ and $\boldsymbol{\Omega}$ whilst the generalized estimating equations are solved to get estimates of $\boldsymbol{\beta}^+$, ϕ, $\boldsymbol{\Omega}$ and also of $\boldsymbol{\alpha}$.

For the latent MRM, let the vector of generalized estimating functions be denoted by $\mathbf{U}^{GEE}$ where $\mathbf{U}^{GEE} = \left(\mathbf{U}_1^{GEE'}, \ \ldots, \mathbf{U}_{J+1}^{GEE'}\right)'$ where

- $\mathbf{U}_j^{GEE}$, $j = 1, \ldots, J$, is the vector of estimating functions involving the derivatives of $m_i\left(\dot{\theta}_i, \boldsymbol{\beta}_j\right)$ with respect to $\beta_{j0}, \ldots, \beta_{jp}$,
- $\mathbf{U}_{J+1}^{GEE}$ is the vector of estimating functions involving the derivatives of $m_i\left(\dot{\theta}_i, \boldsymbol{\beta}_j\right)$ with respect to $\dot{\theta}_1, \ldots, \dot{\theta}_n$.

So as to proceed with deriving the estimating equations for the latent MRM, a working correlation needs to be specified. In general, the choice of this working correlation matrix will influence efficiency of the GEE estimator. The closer the specified correlation matrix to the truth, the higher the efficiency of the GEE estimator. Zeger (1988) has shown that when the correlation between the response variables is not too large, the estimator under the independence working model is relatively efficient. Zhao et al. (1992) concur with Zeger (1988) but also show that 'strong dependencies in the true correlation matrix that are not acknowledged in specifying the working correlation matrix can to lead important loss of efficiency'. Fitzmaurice (1995) shows that efficiency of the resulting estimator may be as low as 60% when compared to the efficiency under the correct correlation structure. A working correlation matrix that may be deemed suitable for time-specific dependency is the first-order autoregressive (AR-1) where elements in the correlation matrix take the form $\mathbb{W}_{jj'} = \alpha^{\left|j-j'\right|}$. If there is no time-specific dependency, it is recommended to use the exchangeable correlation structure, also known as equicorrelation or compound symmetry, for which all off-diagonal elements are equal. When dealing with compositional data, there is invariance in the relabeling of the different components making up the response vector Y_i so the concept of time is irrelevant in this context. A working correlation matrix that does not cater for time-dependency should therefore be used. The exchangeable correlation matrix, a matrix in which all correlations are taken to be equal, or an unstructured correlation matrix, a matrix that assumes unconstrained pairwise correlations where each correlation has to be estimated through the data, are good candidates to be used as working correlation matrices with estimating equations for compositional data. Detail on generalized estimating equations for the latent MRM when an unstructured correlation matrix is used will be provided in the following section. Details for estimating equations with the exchangeable structure follow directly from the work provided for the more general unstructured correlation matrix.

2.4.3.1 Estimating Equations for the Latent MRM under an Unstructured Correlation Matrix

Let the elements of $[\mathbb{W}(\boldsymbol{\alpha})]^{-1}$ be denoted by $\lambda_{jj'}$, $j, j^{'} = 1, \ldots, J$. Under an unstructured correlation matrix, the $(J(p+1)+n)$ generalized estimating equations based on latent $\dot{Y}_{ij}$ are

$$U_{js}^{GEE} = \frac{1}{\phi\sqrt{\omega_j}} \sum_{j'=1}^{J} \left[\frac{\lambda_{jj'}}{\sqrt{\omega_{j'}}} \sum_{i=1}^{n} \left(\frac{\dot{Y}_{ij'}}{m_i\left(\dot{\theta}_i, \boldsymbol{\beta}_{j'}\right)} - 1 \right) x_{i,s-1} \right] = 0 \tag{2.29}$$

$$\text{for } j = 1, \ldots, J, s = 1, \ldots, p+1,$$

$$U_{J+1,i}^{GEE} = \frac{1}{\phi} \sum_{j=1}^{J} \left[\frac{1}{\sqrt{\omega_j}} \left(\frac{\dot{Y}_{ij}}{m_i\left(\dot{\theta}_i, \boldsymbol{\beta}_j\right)} - 1 \right) \sum_{j'=1}^{J} \frac{\lambda_{jj'}}{\sqrt{\omega_{j'}}} \right] = 0 \quad \text{for } i = 1, \ldots, n. \tag{2.30}$$

On using the interchangeability between the latent $\dot{Y}_{ij}$ and compositional Y_{ij}, the estimating equations may alternatively be represented as

$$U_{js}^{GEE} = \frac{1}{\phi\sqrt{\omega_j}} \sum_{j'=1}^{J} \left[\frac{\lambda_{jj'}}{\sqrt{\omega_{j'}}} \sum_{i=1}^{n} \left(\frac{Y_{ij'}}{m_i\left(\theta_i, \boldsymbol{\beta}_{j'}\right)} - 1 \right) x_{i,s-1} \right] = 0 \tag{2.31}$$

$$\text{for } j = 1, \ldots, J, s = 1, \ldots, p+1,$$

$$U_{J+1,i}^{GEE} = \frac{1}{\phi} \sum_{j=1}^{J} \left[\frac{1}{\sqrt{\omega_j}} \left(\frac{Y_{ij}}{m_i\left(\theta_i, \boldsymbol{\beta}_j\right)} - 1 \right) \sum_{j'=1}^{J} \frac{\lambda_{jj'}}{\sqrt{\omega_{j'}}} \right] = 0 \quad \text{for } i = 1, \ldots, n \tag{2.32}$$

where θ_i, $(i = 1, \ldots, n)$, denotes the nuisance parameter attributed to using Y_{ij} instead of $\dot{Y}_{ij}$.

So as to obtain estimates of $\boldsymbol{\beta}$ and $\boldsymbol{\theta}$ from estimating equations (2.31) and (2.32) respectively, an iterative process needs to be used. It is important to note that despite the fact that this section focuses on estimating equations using the unstructured working correlation matrix, the description of a typical GEE estimating procedure which follows, holds for any choice of a working correlation matrix.

An estimation procedure based on that of Liang and Zeger (1986), would involve finding the initial estimate of $\boldsymbol{\beta}^{+} = \left(\boldsymbol{\theta}^{'}, \boldsymbol{\beta}^{'}\right)^{'}$. This is then followed by finding an estimate of $\boldsymbol{\alpha}$ and $\boldsymbol{\Omega}$ given the current estimate of $\boldsymbol{\beta}^{+}$. Once the initial estimates of $\boldsymbol{\alpha}$ and $\boldsymbol{\Omega}$ have been obtained, modified Fisher-scoring is then used to update the estimates of $\boldsymbol{\beta}^{+}$. At every iteration step, $\boldsymbol{\beta}^{+}$ is estimated using the generalized least squares estimator:

$$\widehat{\boldsymbol{\beta}}^{+} = \left(\dot{\mathbb{D}}^{'}\dot{\mathbb{V}}^{-1}\dot{\mathbb{D}}\right)^{-1}\dot{\mathbb{D}}^{'}\dot{\mathbb{V}}^{-1}\mathbf{Z} \tag{2.33}$$

where $\mathbf{Z}$ is the vector of working variates; the latter alternatively known as the modified

dependent variable. In a typical GEE procedure, $\mathbf{Z}$ is worked out using

$$\mathbf{Z} = \dot{\mathbb{D}}\boldsymbol{\beta}^{+} + \mathbf{S} \tag{2.34}$$

where $\mathbf{S}$ is an nJ-vector defined by $\mathbf{S} = \dot{\mathbf{Y}} - \mathbf{m}\left(\boldsymbol{\beta}^{+}\right)$.

Liang and Zeger (1986) refer to the iterative procedure used as modified Fisher scoring due to taking the limiting value (as n tends to infinity) of the expectation of the derivative of the quasi-score functions rather than the expectation of the derivative itself. Also in line with Liang and Zeger (1986), in a typical GEE estimation procedure, estimation of the common dispersion parameter ϕ is carried out once convergence for $\boldsymbol{\beta}^{+}$ is achieved, since ϕ cancels out of the estimating equations (2.31) and (2.32).

The iterative procedure that will be used to estimate the MRM model parameters $\boldsymbol{\beta}^{+}$ will be slightly different. More details on this are presented in the following sections.

2.4.3.2 Performing GLS estimation in Two Steps

In the GEE estimation procedure described in the previous section, the vector of estimates $\widehat{\boldsymbol{\beta}}^{+}$ is estimated through an iterative process where at each iteration, $\boldsymbol{\beta}^{+}$ is estimated using the generalized least squares estimator $\widehat{\boldsymbol{\beta}}^{+} = \left(\dot{\mathbb{D}}'\dot{\mathbb{V}}^{-1}\dot{\mathbb{D}}\right)^{-1}\dot{\mathbb{D}}'\dot{\mathbb{V}}^{-1}\mathbf{Z}$. Alternatively, however, estimation of $\left(\boldsymbol{\theta}', \boldsymbol{\beta}'\right)$ may also be carried out in two different generalized least squares estimation steps. So at each GLS iteration of the estimating procedure, the estimation of $\boldsymbol{\theta}$ is followed by the estimation of $\left(\boldsymbol{\beta}_1', \ldots, \boldsymbol{\beta}_J'\right)$. All these estimates will then be updated during the subsequent iteration. Updating of estimates will continue until convergence, to a specified level of tolerance, is reached.

Suppose that $\widehat{\boldsymbol{\theta}}^{0}$ and $\widehat{\boldsymbol{\beta}}^{0}$ are the initial estimates of $\boldsymbol{\theta}$ and $\boldsymbol{\beta}$ respectively. Also suppose that the estimating functions which are used to obtain $\boldsymbol{\theta}$ and $\boldsymbol{\beta}$ estimates at the $(t+1)^{th}$ iteration are respectively given by $f\left(\boldsymbol{\theta}; \widehat{\boldsymbol{\theta}}^{t}, \widehat{\boldsymbol{\beta}}^{t}\right)$ and $g\left(\boldsymbol{\beta}; \widehat{\boldsymbol{\theta}}^{t+1}, \widehat{\boldsymbol{\beta}}^{t}\right)$. The GLS iterative procedure being described here proceeds in this manner:

- Step 1: Obtain initial estimates $\widehat{\boldsymbol{\theta}}^{0}$ and $\widehat{\boldsymbol{\beta}}^{0}$
- Step 2:
  i. Solve $f\left(\boldsymbol{\theta}; \widehat{\boldsymbol{\theta}}^{t}, \widehat{\boldsymbol{\beta}}^{t}\right) = 0$ to obtain $\widehat{\boldsymbol{\theta}}^{t+1}$
  ii. Solve $g\left(\boldsymbol{\beta}; \widehat{\boldsymbol{\theta}}^{t+1}, \widehat{\boldsymbol{\beta}}^{t}\right) = 0$ to obtain $\widehat{\boldsymbol{\beta}}^{t+1}$
- Repeat step 2 until convergence.

In general, under an unstructured working correlation matrix, we cannot be sure that the two estimating procedures (one-step GLS and two-step GLS) will deliver the same estimates of $\boldsymbol{\beta}$, as we have not investigated the uniqueness of the estimates. However,

as per Wedderburn (1976), under an independence working correlation structure, if estimates arising out of estimating equations for a model with a log link with mean-variance specification for the gamma distribution (as for the latent MRM) exist, they must be unique. For the latent MRM, estimates always exist as they are obtained through the GLS estimator. So under the independence working correlation matrix, estimates of the parameters of the latent MRM are also unique. The uniqueness that is obtained under an independence working correlation is all that really matters in relation to the latent MRM. This is because as shall be seen in the following section, for the latent MRM, the GLS estimator of the parameters of interest $\boldsymbol{\beta}$, is invariant to the values of dispersion and correlation parameters.

2.4.4 Invariance of $\widehat{\boldsymbol{\beta}}$ to the Values of Dispersion and Correlation Parameters

Consider the system of estimating equations based on the general unstructured working correlation matrix, introduced in Section 2.4.3.1. Suppose that updating of the estimates $\widehat{\boldsymbol{\beta}}$ and $\widehat{\dot{\boldsymbol{\theta}}}$ (similarly for $\widehat{\boldsymbol{\theta}}$) is carried out in two separate steps of each iteration. Focus will now be directed towards deriving the expression of the GLS estimator that is used for updating $\widehat{\boldsymbol{\beta}}$ at each iteration. So the linear model (2.34) will be rewritten in terms of $\boldsymbol{\beta}$ rather than in terms of $\boldsymbol{\beta}^{+}$.

Let $\dot{\mathbb{D}}_{\boldsymbol{\beta}}$ denote the $Jn \times J\left(p+1\right)$ matrix of derivatives of $m_i\left(\dot{\theta}_i, \boldsymbol{\beta}_j\right)$ with respect to the β parameters only. The elements in $\dot{\mathbb{D}}_{\boldsymbol{\beta}}$ are given by

$$\frac{\partial m_i\left(\dot{\theta}_i, \boldsymbol{\beta}_j\right)}{\partial \beta_{jk}} = m_i\left(\dot{\theta}_i, \boldsymbol{\beta}_j\right) x_{ik}.$$

For ease of deriving the expression for $\widehat{\boldsymbol{\beta}}$, all the terms in (2.34) will also be divided by the respective means and the resulting terms will be rewritten by component rather than by subject. Thus, the linear model becomes

$$\mathbf{Z}^{*} = \dot{\mathbb{D}}_{\boldsymbol{\beta}}^{*'}\boldsymbol{\beta} + \mathbf{S}^{*},$$

where $\mathbf{Z}^{*} = \left(\mathbf{Z}_{(1)}^{*'}, \ldots, \mathbf{Z}_{(J)}^{*'}\right)'$ and for $i = 1, \ldots, n, j = 1, \ldots, J$, $\mathbf{Z}_{(j)}^{*} = \left(Z_{1j}^{*}, \ldots, Z_{nj}^{*}\right)'$, $Z_{ij}^{*} = Z_{ij}/m_i\left(\dot{\theta}_i, \boldsymbol{\beta}_j\right)$ where Z_{ij} are the elements making up the vector of modified dependent variables $\mathbf{Z}$ in (2.34). Similarly, $\mathbf{S}^{*} = \left(\mathbf{S}_{(1)}^{*'}, \ldots, \mathbf{S}_{(J)}^{*'}\right)'$ where

$$S_{ij}^{*} = \frac{S_{ij}}{m_i\left(\dot{\theta}_i, \boldsymbol{\beta}_j\right)} = \frac{\dot{Y}_{ij}}{m_i\left(\dot{\theta}_i, \boldsymbol{\beta}_j\right)} - 1,$$

and $\dot{\mathbb{D}}^*_{\boldsymbol{\beta}} = \left(\dot{\mathbb{D}}^{*'}_{\boldsymbol{\beta},1}, \ldots, \dot{\mathbb{D}}^{*'}_{\boldsymbol{\beta},J}\right)'$ is a $Jn \times J(p+1)$ matrix which arises out of dividing all the terms in the matrix of derivatives $\dot{\mathbb{D}}_{\boldsymbol{\beta}} = \left(\dot{\mathbb{D}}'_{\boldsymbol{\beta},1}, \ldots, \dot{\mathbb{D}}'_{\boldsymbol{\beta},J}\right)'$ by the respective means giving

$$\dot{\mathbb{D}}^*_{\boldsymbol{\beta}} = \mathbb{I}_J \otimes \mathbb{X}, \tag{2.35}$$

where for $j = 1, \ldots, J$, $\dot{\mathbb{D}}_{\boldsymbol{\beta},j}$ denotes the matrix of derivatives of $m_i\left(\dot{\theta}_i, \boldsymbol{\beta}_j\right)$ with respect to the β parameters corresponding to the j^{th} component, $\mathbb{I}_J$ is a $J \times J$ identity matrix, $\otimes$ is the Kronecker product and $\mathbb{X}$ is the design matrix.

Furthermore, from Section 2.2,

$$\left(\frac{\dot{Y}_{i1}}{m_i\left(\dot{\theta}_i, \boldsymbol{\beta}_1\right)}, \ldots, \frac{\dot{Y}_{iJ}}{m_i\left(\dot{\theta}_i, \boldsymbol{\beta}_J\right)}\right)' = \mathbf{E}_i$$

and $\text{Var}(\mathbf{E}_i) = \dot{\boldsymbol{\Sigma}}$.

Let $\dot{\mathbb{V}}^*$ denote the variance-covariance matrix for all error terms E_{ij}. The matrix $\dot{\mathbb{V}}^*$ is equivalently the variance-covariance matrix of all the elements making up $\mathbf{S}^*$. Since $\dot{\boldsymbol{\Sigma}}$ is assumed to be the same for all i, $\dot{\mathbb{V}}^*$ is given by

$$\dot{\mathbb{V}}^* = \dot{\boldsymbol{\Sigma}} \otimes \mathbb{I}_n \tag{2.36}$$

giving the GLS weight matrix $\dot{\mathbb{V}}^{*-1} = \dot{\boldsymbol{\Sigma}}^{-1} \otimes \mathbb{I}_n$, where $\mathbb{I}_n$ is an $n \times n$ identity matrix.

It then follows that

$$\begin{aligned}
\left(\dot{\mathbb{D}}^{*'}_{\boldsymbol{\beta}} \dot{\mathbb{V}}^{*-1} \dot{\mathbb{D}}^*_{\boldsymbol{\beta}}\right)^{-1} &= \left(\left[\mathbb{I}_J \otimes \mathbb{X}'\right]\left[\dot{\boldsymbol{\Sigma}}^{-1} \otimes \mathbb{I}_n\right]\left[\mathbb{I}_J \otimes \mathbb{X}\right]\right)^{-1} \\
&= \left[\dot{\boldsymbol{\Sigma}}^{-1} \otimes \mathbb{X}'\mathbb{X}\right]^{-1} \\
&= \dot{\boldsymbol{\Sigma}} \otimes \left(\mathbb{X}'\mathbb{X}\right)^{-1},
\end{aligned}$$

so that the GLS estimator is given by

$$\begin{aligned}
\widehat{\boldsymbol{\beta}} &= \left(\dot{\mathbb{D}}^{*'}_{\boldsymbol{\beta}} \dot{\mathbb{V}}^{*-1} \dot{\mathbb{D}}^*_{\boldsymbol{\beta}}\right)^{-1} \dot{\mathbb{D}}^{*'}_{\boldsymbol{\beta}} \dot{\mathbb{V}}^{*-1} \mathbf{Z}^* \\
&= \left[\dot{\boldsymbol{\Sigma}} \otimes \left(\mathbb{X}'\mathbb{X}\right)^{-1}\right]\left[\mathbb{I}_J \otimes \mathbb{X}'\right]\left[\dot{\boldsymbol{\Sigma}}^{-1} \otimes \mathbb{I}_n\right]\mathbf{Z}^* \\
&= \left[\mathbb{I}_J \otimes \left(\mathbb{X}'\mathbb{X}\right)^{-1}\mathbb{X}'\right]\mathbf{Z}^*
\end{aligned} \tag{2.37}$$

showing that the GLS estimator $\widehat{\boldsymbol{\beta}}$ is invariant to the values of the correlation and dispersion parameters. The derivation for the above result is motivated by the familiar invariance

property of the GLS estimator in multivariate linear regression (e.g. Mardia et al., 1979, p. 173).

The invariance of the GLS estimator (2.37) is a very appealing property of this system of estimating equations as on considering equal dispersion parameters and an independence working correlation matrix, the generalized estimating equations reduce to quasi-likelihood equations (2.17) and (2.18) with equal dispersion parameters. On using this system, however, there is an issue which deserves attention.

From equation (2.7), $\sum_{j=1}^{J} \dot{Y}_{ij} = c_i$. It follows that $\sum_{j=1}^{J} E\left(\dot{Y}_{ij}\right) = \sum_{j=1}^{J} m_i\left(\dot{\theta}_i, \boldsymbol{\beta}_j\right) = c_i$. Due to the sum constraint on the means of each case i, once estimation of the model parameters is carried out, it is desirable to obtain fitted values that also sum to c_i for case i. This rescaling of the fitted values may indeed be carried out once the resulting estimates of $\boldsymbol{\beta}$ have reached convergence in the iterative process. Alternatively, rescaling of the fitted values may also be carried out throughout the iterative process. This alternative strategy may be implemented by setting up a new set of estimating equations for $\dot{\boldsymbol{\theta}}$ whilst keeping the same estimating equations for $\boldsymbol{\beta}$. From now onwards, the previous system of estimating equations will be referred to by the name *standard* system. The new system of estimating equations, to be introduced in the following section, will be referred to by the name *hybrid* system.

2.4.5 A Hybrid System of Estimating Equations

The hybrid system being introduced here is an alternative system of estimating equations, which caters for the rescaling of fitted values throughout the iterative process. The two-step procedure explained in Section 2.4.3.2 still holds here. The only difference now is that a different set of estimating equations is used to estimate $\dot{\boldsymbol{\theta}}$. More specifically, once estimates of $\boldsymbol{\beta}$ have been obtained through estimating equations (2.17) with equal dispersion parameters, using generalized least squares estimation, the estimator of $\dot{\theta}_i$ will be that which satisfies $\sum_{j=1}^{J} m_i\left(\dot{\theta}_i, \widehat{\boldsymbol{\beta}}_j\left(\widehat{\dot{\theta}}_i\right)\right) = c_i$. Let $U^H_{J+1,i}$ denote the estimating function used to solve for $\dot{\theta}_i$. Then, under the hybrid system, the estimate $\widehat{\dot{\theta}}_i$ is obtained by solving

$$U^H_{J+1,i} = \sum_{j=1}^{J}\left[\dot{Y}_{ij} - m_i\left(\dot{\theta}_i, \widehat{\boldsymbol{\beta}}_j\left(\widehat{\dot{\theta}}_i\right)\right)\right] = 0. \tag{2.38}$$

Since in practice it is the compositional response variables Y_{ij} that will be available and because it has been shown that a change in the value of c_i in (2.6) will lead to the same estimates of $\boldsymbol{\beta}$, the value of c_i will be taken to be equal to one even here. Equation (2.38) may thus be rewritten in terms of Y_{ij} by replacing $\dot{Y}_{ij}$ with Y_{ij} giving

$$U^{H}_{J+1,i} = \sum_{j=1}^{J} \left[Y_{ij} - m_i \left(\theta_i, \widehat{\boldsymbol{\beta}}_j \left(\widehat{\theta}_i \right) \right) \right] = 0. \tag{2.39}$$

Seeing that under the standard and the hybrid systems of estimating equations, the β parameters are estimated using the same estimating equations, the GLS expression used to obtain estimates of the β parameters for the standard system, also holds under the hybrid system.

Now due to the way that the hybrid system is set up, we have seen that once the iterative process is ready, the estimated means for each case i sum to 1, which is indeed a desirable property in dealing with compositional data. There is another property which makes the hybrid system stand out even more. In the section which follows, we shall see how the hybrid system of estimating equations for $J = 2$ is in fact equivalent to the system of equations used by Wedderburn (1974) in his study on barley leaf data. In Section 2.6, we will then develop a new method which generalizes the equivalence of the hybrid system to estimating equations obtained through a multivariate type of logit model to the case where $J > 2$. All the just mentioned makes the hybrid system a better choice, for estimating the parameters of the latent MRM, over the standard system.

2.5 The Equivalence of the Hybrid Estimating Equations to Wedderburn's Estimating Equations when $J = 2$

Due to the fact that $\widehat{\boldsymbol{\beta}}$ has been shown to be invariant to the values of dispersion and correlation parameters, the hybrid system assumes equal dispersion parameters and independence between $\dot{Y}_{ij}$s. The system of estimating equations being considered will also be that of an overparametrized system, where $(\theta_1, \ldots, \theta_n)$ are estimated through (2.39) and $(\beta_{10}, \ldots, \beta_{1p}, \ldots, \beta_{J0}, \ldots, \beta_{Jp})$ are estimated using

$$U_{js} = \sum_{i=1}^{n} \left(\frac{Y_{ij}}{m_i \left(\theta_i, \boldsymbol{\beta}_j \right)} - 1 \right) x_{i,s-1} = 0 \tag{2.40}$$

for $j = 1, \ldots, J$.

Since $\sum_{j=1}^{J} Y_{ij} = 1$, rearranging (2.39) leads to

$$\exp\left(\theta_i\right) = \frac{1}{\sum_{j=1}^{J} \exp\left(\mathbf{x}_i' \boldsymbol{\beta}_j \right)}. \tag{2.41}$$

Substituting (2.41) in (2.40) gives

$$\sum_{i=1}^{n}\left(\frac{Y_{ij}}{p_{ij}}-1\right)x_{i,s-1}=0, \tag{2.42}$$

where

$$p_{ij}=\frac{\exp\left(\mathbf{x}_i^{'}\boldsymbol{\beta}_j\right)}{\sum_{j^{'}=1}^{J}\exp\left(\mathbf{x}_i^{'}\boldsymbol{\beta}_{j^{'}}\right)}. \tag{2.43}$$

Now, for $J=2$, the parameter space of interest under an overparametrized system is made up of $\left(\boldsymbol{\beta}_1^{'},\boldsymbol{\beta}_2^{'}\right)^{'}$. Due to the system being overparametrized, identification of all parameters $(\beta_{10},\ldots,\beta_{2p})$, will not be possible. Consider once again the argument on identification posed on Pg 26 and apply it for $J=2$. Since identification and thus also interpretation is only possible for a set of contrasts for the parameters, in this case we will take the last (second) component as reference component and we will focus on the difference $(\boldsymbol{\beta}_1-\boldsymbol{\beta}_2)$. Call this difference $\boldsymbol{\gamma}$. The estimating equations for $\boldsymbol{\gamma}$ under the reparametrization are given by

$$\sum_{i=1}^{n}\left[\left(\frac{Y_{i1}}{p_{i1}}-1\right)-\left(\frac{Y_{i2}}{p_{i2}}-1\right)\right]x_{i,s-1}=0. \tag{2.44}$$

Now $p_{i1}+p_{i2}=1$ and $Y_{i1}+Y_{i2}=1$, so substituting for $p_{i2}=1-p_{i1}$, and $Y_{i2}=1-Y_{i1}$ in (2.44) gives

$$\sum_{i=1}^{n}\left(\frac{Y_{i1}-p_{i1}}{p_{i1}\left(1-p_{i1}\right)}\right)x_{i,s-1}=0 \tag{2.45}$$

where

$$p_{i1}=\frac{\exp\left(\mathbf{x}_i^{'}\boldsymbol{\gamma}\right)}{1+\exp\left(\mathbf{x}_i^{'}\boldsymbol{\gamma}\right)}. \tag{2.46}$$

The resulting system of estimating equations is the same as that achieved by Wedderburn (1974) for the analysis of barley leaf data, in which a logit link function is used together with a mean-variance relationship defined by $V(p_{i1})=p_{i1}^2\left(1-p_{i1}\right)^2$.

Application to the Barley Leaf Data

The hybrid system of estimating equations with $J=2$ and Wedderburn's (1974) logit model have been fitted to the barley leaf data analyzed in Wedderburn (1974). The barley leaf data, available in the *gnm* package in R (R Core Team, 2016), gives the percentage of leaf blotch found on 10 varieties of barley, grown at 9 different sites. So as to use the hybrid system of estimating equations, two compositional response variables have been created, one variable, Y_1, containing the percentage 'With Leaf Blotch' and the other variable, Y_2, containing the percentage 'Without Leaf Blotch'. Variety and Site are considered as the

explanatory variables X_1 and X_2 respectively. The hybrid system of estimating equations has been implemented using the newly developed *cglm* package. For a brief introduction on the *cglm* package see Chapter 5. Wedderburn's logit model has been fitted to the data using family specification wedderburn in the *glm* function available in R. As expected, the estimates obtained using the hybrid system and Wedderburn's logit model are the same. The two resulting sets of estimates match up to seven decimal places using a tolerance of 10^{-14}. The estimates are provided in Table 2.1. In Section 2.8.2, it will also be shown how the standard errors obtained through the hybrid system with $J = 2$ are also equivalent to those obtained using Wedderburn's logit model.

	Estimate
(Intercept)	-7.9224
siteB	1.3831
siteC	3.8601
siteD	3.5570
siteE	4.1079
siteF	4.3054
siteG	4.9181
siteH	5.6949
siteI	7.0676
variety2	-0.4674
variety3	0.0788
variety4	0.9541
variety5	1.3526
variety6	1.3285
variety7	2.3401
variety8	3.2626
variety9	3.1355
variety10	3.8873

Table 2.1: Barley Leaf Data Parameter Estimates

The relationship between the hybrid system and Wedderburn's system of estimating equations makes the hybrid system very appealing, as the γ estimates achieved under the hybrid system will inherit all the desirable properties of Wedderburn's quasi-likelihood estimator. Indeed though, this equivalence has so far been shown to hold only for the case $J = 2$. It would be ideal if such an equivalence could be generalized to cater for the case where $J > 2$. In the section which follows, a new method which generalizes the equivalence of the hybrid system to estimating equations obtained through a multivariate type of logit model to the case where $J > 2$ will be developed.

2.6 Extending Wedderburn's Estimating Equations to the Case where J is Greater than 2

In view of the role played by the overparametrized system in relating the hybrid system of estimating equations (2.42), with $J = 2$, and Wedderburn's estimating equations, a generalization of this relationship will be sought by once again considering an overparametrized system for $J > 2$.

If an overparametrized system is considered, identification of all the β parameters of the latent MRM will not be possible. As has been done for the case $J = 2$, instead of considering the hybrid system of equations (2.42) for the β parameters, the estimating equations will be modified to estimate the difference in parameters. Taking the last component as reference component, without loss of generality, the differences under consideration are $\left(\left(\boldsymbol{\beta}_1 - \boldsymbol{\beta}_J\right), \ldots, \left(\boldsymbol{\beta}_{J-1} - \boldsymbol{\beta}_J\right)\right)$. Call these differences $\left(\boldsymbol{\gamma}_1, \ldots, \boldsymbol{\gamma}_{J-1}\right)$.

By considering differences in parameters, with the last component taken as reference, the multinomial logistic model

$$\begin{aligned} \eta_{ij} &= \log\left(p_{ij}\right) - \log\left(p_{iJ}\right) \\ &= \left(\beta_{j0} - \beta_{J0}\right) x_{i0} + \cdots + \left(\beta jp - \beta_{Jp}\right) x_{ip} \\ &= \gamma_{j0} x_{i0} + \cdots + \gamma_{jp} x_{ip} \end{aligned} \tag{2.47}$$

is being specified since

$$p_{ij} = \frac{\exp\left(\mathbf{x}_i^{'}\boldsymbol{\beta}_j\right)}{\sum_{j'=1}^{J} \exp\left(\mathbf{x}_i^{'}\boldsymbol{\beta}_{j'}\right)} = \frac{\exp\left(\mathbf{x}_i^{'}\left(\boldsymbol{\beta}_j - \boldsymbol{\beta}_J\right)\right)}{\sum_{j'=1}^{J} \exp\left(\mathbf{x}_i^{'}\left(\boldsymbol{\beta}_{j'} - \boldsymbol{\beta}_J\right)\right)} = \frac{\exp\left(\mathbf{x}_i^{'}\boldsymbol{\gamma}_j\right)}{\sum_{j'=1}^{J} \exp\left(\mathbf{x}_i^{'}\boldsymbol{\gamma}_{j'}\right)}. \tag{2.48}$$

The matrix of derivatives, $\mathbb{D}_i = \frac{\partial \mathbf{p}_i}{\partial \boldsymbol{\gamma}}$, under such a model, is given by

$$\mathbb{D}_i = \frac{\partial \mathbf{p}_i}{\partial \boldsymbol{\eta}_i}\mathbb{X}_i = \left(\mathbb{P}_i - \mathbf{p}_i\mathbf{p}_i^{'}\right)\mathbb{X}_i \tag{2.49}$$

where $\mathbf{p}_i$ is a J-vector of proportions $(p_{i1}, \ldots, p_{iJ})$, $\boldsymbol{\gamma} = \left(\boldsymbol{\gamma}_1^{'}, \ldots, \boldsymbol{\gamma}_{J-1}^{'}\right)^{'}$, $\boldsymbol{\eta}_i = \left(\eta_{i1}, \ldots, \eta_{iJ}\right)^{'}$, $\mathbb{P}_i = \operatorname{diag}\left(p_{i1}, \ldots, p_{iJ}\right)$ and $\mathbb{X}_i$ is a $J \times (J-1)(p+1)$ defined by

$$\mathbb{X}_i = \begin{pmatrix} \mathbb{I}_{J-1} \otimes \left(x_{i0} \ldots x_{ip}\right) \\ \mathbf{0}^{'} \end{pmatrix}.$$

Let $\mathbf{U} = \left(\mathbf{U}_1^{'}, \ldots, \mathbf{U}_{J-1}^{'}\right)^{'}$, where $\mathbf{U}_j$ is a $(p+1)$-vector that involves the derivatives of p_{ij} with respect to $(\gamma_{j0}, \ldots, \gamma_{jp})$, for $j = 1, \ldots, J-1$.

Having identified the structure of $\mathbb{D}_i$, in order to construct a system of estimating equations for J-compositional response variables, the aim now is that of identifying the implied form of the variance-covariance matrix of $\mathbf{Y}_i$, call this matrix $\mathbb{V}_i$, such that as per McCullagh (1983), the system of estimating equations that may be used to estimate the parameters $(\boldsymbol{\gamma}_1, \ldots, \boldsymbol{\gamma}_{J-1})$ takes the form

$$\mathbf{U} = \frac{1}{\phi}\sum_{i=1}^{n} \mathbb{D}_i' \mathbb{V}_i^{-} (\mathbf{Y}_i - \mathbf{p}_i) = \frac{1}{\phi}\sum_{i=1}^{n} \mathbb{X}_i' \left(\mathbb{P}_i - \mathbf{p}_i \mathbf{p}_i^{'}\right) \mathbb{V}_i^{-} (\mathbf{Y}_i - \mathbf{p}_i) = \mathbf{0}, \tag{2.50}$$

where a generalized inverse is invoked for $\mathbb{V}_i$ due to this variance-covariance matrix being a singular matrix.

Now on comparing (2.50) with the hybrid system (2.42), we know that

$$\left(\mathbb{P}_i - \mathbf{p}_i \mathbf{p}_i^{'}\right) \mathbb{V}_i^{-} = \mathbb{P}_i^{-1}, \tag{2.51}$$

so $\mathbb{V}_i^{-} = \left[\mathbb{P}_i - \mathbf{p}_i \mathbf{p}_i^{'}\right]^{+} \mathbb{P}_i^{-1}$, where $\left[\mathbb{P}_i - \mathbf{p}_i \mathbf{p}_i^{'}\right]^{+}$ is the Moore-Penrose pseudoinverse of $\mathbb{P}_i - \mathbf{p}_i \mathbf{p}_i^{'}$ and as per Tanabe and Sagae (1992),

$$\left(\mathbb{P}_i - \mathbf{p}_i \mathbf{p}_i^{'}\right)^{+} = \left(\mathbb{I}_J - \frac{1}{J}\mathbf{1}\mathbf{1}^{'}\right) \mathbb{P}_i^{-1} \left(\mathbb{I}_J - \frac{1}{J}\mathbf{1}\mathbf{1}^{'}\right), \tag{2.52}$$

where $\mathbb{I}_J$ is a $J \times J$ identity matrix, $\mathbf{1}$ is a J-vector of ones and $\mathbf{1}\mathbf{1}^{'}$ is a $J \times J$ matrix of ones.

It therefore follows that

$$\mathbb{V}_i^{-} = \left(\mathbb{I}_J - \frac{1}{J}\mathbf{1}\mathbf{1}^{'}\right) \mathbb{P}_i^{-1} \left(\mathbb{I}_J - \frac{1}{J}\mathbf{1}\mathbf{1}^{'}\right) \mathbb{P}_i^{-1}. \tag{2.53}$$

Substituting (2.53) in (2.50), yields the estimating equations to be used with J-compositional response variables

$$\mathbf{U} = \frac{1}{\phi}\sum_{i=1}^{n} \mathbb{X}_i' \left(\mathbb{P}_i - \mathbf{p}_i \mathbf{p}_i^{'}\right) \mathbb{V}_i^{-} (\mathbf{Y}_i - \mathbf{p}_i) = \frac{1}{\phi}\sum_{i=1}^{n} \mathbb{X}_i' \left(\mathbb{I}_J - \frac{1}{J}\mathbf{1}\mathbf{1}^{'}\right) \mathbb{P}_i^{-1} (\mathbf{Y}_i - \mathbf{p}_i) = \mathbf{0}, \tag{2.54}$$

since $\left(\mathbb{P}_i - \mathbf{p}_i \mathbf{p}_i^{'}\right) \left(\mathbb{I}_J - \frac{1}{J}\mathbf{1}\mathbf{1}^{'}\right) \mathbb{P}_i^{-1} = \left(\mathbb{I}_J - \frac{1}{J}\mathbf{1}\mathbf{1}^{'}\right)$ and $\mathbb{I}_J - \frac{1}{J}\mathbf{1}\mathbf{1}^{'}$ is idempotent.

For $j = 1, \ldots, J-1$ and $s = 0, \ldots, p$, the equations making up (2.54) may be shown to be in the form

$$U_{js} = \frac{1}{\phi}\sum_{i=1}^{n} \left(\frac{Y_{ij}}{p_{ij}} - \frac{1}{J}\sum_{j'=1}^{J} \frac{Y_{ij'}}{p_{ij'}}\right) x_{is} = 0. \tag{2.55}$$

In what follows, the term p_{ij} in estimating equations (2.55) (or equivalently (2.54)), will be considered to be expressed as

$$p_{ij} = \frac{\exp\left(\mathbf{x}_i^{'}\boldsymbol{\gamma}_j\right)}{\sum_{j'=1}^{J}\exp\left(\mathbf{x}_i^{'}\boldsymbol{\gamma}_{j'}\right)}. \tag{2.56}$$

Also, from now onwards, estimating equations for compositional data (2.54) (or equivalently (2.55)), will be referred to by the name *generalized Wedderburn* estimating equations.

2.7 A Working Variance-Covariance Structure for Compositional Response Variables

In this section we will show that the mean-model specified under the generalized Wedderburn approach is related to the first order Taylor Series approximation of the mean of $\mathbf{Y}_i$, and we will also define the working variance-covariance structure $\mathbb{V}_i$ for compositional response variables by using the first order Taylor Series approximation of the variance of Y_{ij}. The first order Taylor Series approximation to the mean and the variance of Y_{ij} are presented first. For ease of notation, in the lemma and theorem which follow, the notation $m_i\left(\theta_i, \boldsymbol{\beta}_j\right)$ is replaced by m_{ij}.

Lemma 2.7.1. *Consider the vector of latent variables* $\dot{\mathbf{Y}}_i = \left(\dot{Y}_{i1}, \ldots, \dot{Y}_{iJ}\right)^{'}$, $i = 1, \ldots, n$. *Let* $\mathbf{Y}_i$ *be the corresponding vector of compositional variables so that its components are given by* $Y_{ij} = \frac{\dot{Y}_{ij}}{\dot{Y}_{i1}+\cdots+\dot{Y}_{iJ}}$, $j = 1, \ldots, J$. *Suppose that* $E\left(\dot{\mathbf{Y}}_i\right) = \mathbf{m}_i$ *with components* $(m_{i1}, \ldots, m_{iJ})$, *and that* $\text{Var}\left(\dot{\mathbf{Y}}_i\right) = \mathbb{A}_i^{\frac{1}{2}}\dot{\boldsymbol{\Sigma}}\mathbb{A}_i^{\frac{1}{2}}$, *where* $\mathbb{A}_i$ *is a diagonal matrix with elements* $\left(m_{i1}^2, \ldots, m_{iJ}^2\right)$ *and* $\dot{\boldsymbol{\Sigma}}$ *is as defined in (2.4). Also, let* $\mathbf{p}_i = (p_{i1}, \ldots, p_{iJ})'$ *where* $p_{ij} = \frac{m_{ij}}{m_{i1}+\cdots+m_{iJ}}$ *and* $\mathbb{P}_i = \text{diag}\left(p_{i1}, \ldots, p_{iJ}\right)$. *Using a first order Taylor Series approximation of* $\mathbf{h}\left(\dot{\mathbf{y}}_i\right) = \mathbf{y}_i$ *around its mean vector* $\mathbf{h}\left(\mathbf{m}_i\right)$ *gives*

$$E\left(\mathbf{Y}_i\right) = E\left(\mathbf{h}\left(\dot{\mathbf{Y}}_i\right)\right) \approx \mathbf{p}_i \tag{2.57}$$

and

$$\text{Var}\left(\mathbf{Y}_i\right) = \text{Var}\left(\mathbf{h}\left(\dot{\mathbf{Y}}_i\right)\right) \approx \left(\mathbb{P}_i - \mathbf{p}_i\mathbf{p}_i^{'}\right)\dot{\boldsymbol{\Sigma}}\left(\mathbb{P}_i - \mathbf{p}_i\mathbf{p}_i^{'}\right). \tag{2.58}$$

Proof. The first order Taylor Series approximation of $\mathbf{h}\left(\dot{\mathbf{y}}_i\right)$ around its mean vector $\mathbf{h}\left(\mathbf{m}_i\right)$ is given by

$$\mathbf{h}\left(\dot{\mathbf{y}}_i\right) \approx \mathbf{h}\left(\mathbf{m}_i\right) + \nabla\mathbf{h}\left(\mathbf{m}_i\right)\left(\dot{\mathbf{y}}_i - \mathbf{m}_i\right), \tag{2.59}$$

where $\nabla \mathbf{h}\left(\mathbf{m}_i\right) = \left[\frac{\partial \mathbf{h}(\dot{\mathbf{y}}_i)}{\partial \dot{\mathbf{y}}_i}\right]_{\dot{\mathbf{y}}_i = \mathbf{m}_i}$. Thus

$$E\left(\mathbf{h}\left(\dot{\mathbf{Y}}_i\right)\right) \approx \mathbf{h}\left(\mathbf{m}_i\right) = \left(\frac{m_{i1}}{m_{i1}+\cdots+m_{iJ}}, \ldots, \frac{m_{iJ}}{m_{i1}+\cdots+m_{iJ}}\right)' = \left(p_{i1}, \ldots, p_{iJ}\right)' = \mathbf{p}_i$$

and

$$\text{Var}\left(\mathbf{h}\left(\dot{\mathbf{Y}}_i\right)\right) \approx \nabla \mathbf{h}\left(\mathbf{m}_i\right) \text{Var}\left(\dot{\mathbf{Y}}_i\right) \left(\nabla \mathbf{h}\left(\mathbf{m}_i\right)\right)'.$$

Now $\frac{\partial \mathbf{h}(\dot{\mathbf{y}}_i)}{\partial \dot{\mathbf{y}}_i} = \frac{1}{\dot{y}_{i1}+\cdots+\dot{y}_{iJ}}\left(\mathbb{I}_J - \frac{1}{\dot{y}_{i1}+\cdots+\dot{y}_{iJ}}\dot{\mathbf{y}}_i \mathbf{1}'\right)$, so

$$\begin{aligned}
\nabla \mathbf{h}\left(\mathbf{m}_i\right) = \left[\frac{\partial \mathbf{h}\left(\dot{\mathbf{y}}_i\right)}{\partial \dot{\mathbf{y}}_i}\right]_{\dot{\mathbf{y}}_i = \mathbf{m}_i} &= \frac{1}{m_{i1}+\cdots+m_{iJ}}\left(\mathbb{I}_J - \frac{1}{m_{i1}+\cdots+m_{iJ}}\mathbf{m}_i \mathbf{1}'\right) \\
&= \frac{1}{m_{i1}+\cdots+m_{iJ}}\left(\mathbb{I}_J - \mathbf{p}_i \mathbf{1}'\right).
\end{aligned}$$

Thus

$$\begin{aligned}
\text{Var}\left(\mathbf{h}\left(\dot{\mathbf{Y}}_i\right)\right) &\approx \left(\frac{1}{m_{i1}+\cdots+m_{iJ}}\right)^2 \left(\mathbb{I}_J - \mathbf{p}_i \mathbf{1}'\right) \text{Var}\left(\dot{\mathbf{Y}}_i\right) \left(\mathbb{I}_J - \mathbf{p}_i \mathbf{1}'\right) \\
&= \left(\mathbb{I}_J - \mathbf{p}_i \mathbf{1}'\right) \mathbb{P}_i \dot{\mathbf{\Sigma}} \mathbb{P}_i \left(\mathbb{I}_J - \mathbf{p}_i \mathbf{1}'\right) \\
&= \left(\mathbb{P}_i - \mathbf{p}_i \mathbf{p}_i{}'\right) \dot{\mathbf{\Sigma}} \left(\mathbb{P}_i - \mathbf{p}_i \mathbf{p}_i{}'\right).
\end{aligned}$$

□

From the Taylor Series approximation we can see that $E\left(Y_{ij}\right) = p_{ij} + O\left(\phi\right)$. Since the estimating equations (2.54), used under the generalized Wedderburn approach, are in terms of p_{ij}, this shows that the mean-model specified under the generalized Wedderburn approach ($E\left(Y_{ij}\right) = p_{ij}$) will only be equal to the actual mean of Y_{ij} when the dispersion parameter ϕ is negligible. Since there will always be some level of dispersion in the model, the estimates that result through the generalized Wedderburn model will be different from the estimates that would result from the multiplicative model (2.1).

By means of the first order Taylor Series approximation we also get that $\text{Var}\left(\mathbf{Y}_i\right) \approx \left(\mathbb{P}_i - \mathbf{p}_i \mathbf{p}_i{}'\right) \dot{\mathbf{\Sigma}} \left(\mathbb{P}_i - \mathbf{p}_i \mathbf{p}_i{}'\right)$. Once again, the smaller the dispersion in the model, the more precise is the first order Taylor Series approximation to the true variance-covariance structure of $\mathbf{Y}_i$. In general, there will always be some level of dispersion in the model so the true $\text{Var}\left(\mathbf{Y}_i\right)$ might be approximately equal to

$$\left(\mathbb{P}_i - \mathbf{p}_i \mathbf{p}_i{}'\right) \mathbf{\Sigma} \left(\mathbb{P}_i - \mathbf{p}_i \mathbf{p}_i{}'\right), \tag{2.60}$$

where $\mathbf{\Sigma}$ is some variance-covariance matrix.

Using (2.60), we thus define the working variance-covariance matrix of $\mathbf{Y}_i$, to be used in the generalized Wedderburn system of estimating equations for compositional variables,

to be

$$\phi\mathbb{V}_{\mathbf{p}_i,\boldsymbol{\Omega},\mathbb{W}} = \left(\mathbb{P}_i - \mathbf{p}_i\mathbf{p}_i^{'}\right)\boldsymbol{\Sigma}\left(\mathbb{P}_i - \mathbf{p}_i\mathbf{p}_i^{'}\right), \tag{2.61}$$

where we write $\boldsymbol{\Sigma}$ in the form

$$\boldsymbol{\Sigma} = \phi\boldsymbol{\Omega}^{\frac{1}{2}}\mathbb{W}\boldsymbol{\Omega}^{\frac{1}{2}}, \tag{2.62}$$

with ϕ being a common dispersion parameter, $\boldsymbol{\Omega} = \text{diag}\,(\omega_1, \ldots, \omega_J)$ with $\omega_1, \ldots, \omega_J$ being relative dispersion parameters and $\mathbb{W}$ a correlation matrix. Note that the parameters ϕ, $\boldsymbol{\Omega}$ and $\mathbb{W}$ used in (2.62) are different from those used in (2.4), but the same notation is being used here for simplicity.

Now, estimating equations (2.54) are solved to get an estimate of $\boldsymbol{\gamma}_j$, for $j = 1, \ldots, J-1$. The parameter $\boldsymbol{\gamma}_j$ is $\boldsymbol{\beta}_j - \boldsymbol{\beta}_J$ and in Section 2.4.4 it has been shown that the $\widehat{\boldsymbol{\beta}}_j$ is invariant to the values of dispersion and correlation parameters. So in solving estimating equations (2.54), for ease of computation, we can actually use $\mathbb{V}_{\mathbf{p}_i,\mathbb{I}_J,\mathbb{I}_J}$ instead of $\mathbb{V}_{\mathbf{p}_i,\boldsymbol{\Omega},\mathbb{W}}$, where

$$\mathbb{V}_{\mathbf{p}_i,\mathbb{I}_J,\mathbb{I}_J} = \left(\mathbb{P}_i - \mathbf{p}_i\mathbf{p}_i^{'}\right)\left(\mathbb{P}_i - \mathbf{p}_i\mathbf{p}_i^{'}\right). \tag{2.63}$$

In the theorem which follows, $\mathbb{V}_i^-$ of equation (2.53) will in fact be shown to be a generalized inverse of $\mathbb{V}_{\mathbf{p}_i,\mathbb{I}_J,\mathbb{I}_J}$.

Theorem 2.7.1. *Consider the latent variables $\dot{Y}_{ij}$, $(i = 1, \ldots, n,\ j = 1, \ldots, J)$. Let Y_{ij} be the corresponding compositional variables $Y_{ij} = \frac{\dot{Y}_{ij}}{\dot{Y}_{i1}+\cdots+\dot{Y}_{iJ}}$ and suppose that $E\left(\dot{Y}_{ij}\right) = m_{ij}$ and that $p_{ij} = \frac{m_{ij}}{m_{i1}+\cdots+m_{iJ}}$. Then, $\mathbb{V}_i^-$ is a generalized inverse of $\mathbb{V}_{\mathbf{p}_i,\mathbb{I}_J,\mathbb{I}_J} = \left(\mathbb{P}_i - \mathbf{p}_i\mathbf{p}_i^{'}\right)$ $\times\left(\mathbb{P}_i - \mathbf{p}_i\mathbf{p}_i^{'}\right)$, where $\mathbb{V}_i^-$ is as defined in (2.53), $\mathbb{P}_i = \text{diag}\,(p_{i1}, \ldots, p_{iJ})$ and $\mathbf{p}_i$ is the J-vector of proportions defined by $\mathbf{p}_i = (p_{i1}, \ldots, p_{iJ})^{'}$.*

Proof. To show that $\mathbb{V}_i^-$ is a generalized inverse of $\mathbb{V}_{\mathbf{p}_i,\mathbb{I}_J,\mathbb{I}_J}$, it is required to show that

$$\mathbb{V}_{\mathbf{p}_i,\mathbb{I}_J,\mathbb{I}_J} = \mathbb{V}_{\mathbf{p}_i,\mathbb{I}_J,\mathbb{I}_J}\mathbb{V}_i^-\mathbb{V}_{\mathbf{p}_i,\mathbb{I}_J,\mathbb{I}_J}.$$

Using (2.53),

$$\mathbb{V}_i^- = \left(\mathbb{I}_J - \frac{1}{J}\mathbf{1}\mathbf{1}^{'}\right)\mathbb{P}_i^{-1}\left(\mathbb{I}_J - \frac{1}{J}\mathbf{1}\mathbf{1}^{'}\right)\mathbb{P}_i^{-1},$$

where $\mathbb{I}_J$ is a $J \times J$ identity matrix and $\mathbf{1}\mathbf{1}^{'}$ is a $J \times J$ matrix of ones.

Then, since

$$\left(\mathbb{P}_i - \mathbf{p}_i\mathbf{p}_i^{'}\right)\left(\mathbb{I}_J - \frac{1}{J}\mathbf{1}\mathbf{1}^{'}\right) = \mathbb{P}_i - \mathbf{p}_i\mathbf{p}_i^{'},$$

and

$$\left(\mathbb{I}_J - \frac{1}{J}\mathbf{1}\mathbf{1}^{'}\right)\mathbb{P}_i^{-1}\left(\mathbb{P}_i - \mathbf{p}_i\mathbf{p}_i^{'}\right) = \left[\left(\mathbb{P}_i - \mathbf{p}_i\mathbf{p}_i^{'}\right)\mathbb{P}_i^{-1}\left(\mathbb{I}_J - \frac{1}{J}\mathbf{1}\mathbf{1}^{'}\right)\right]^{'} = \mathbb{I}_J - \frac{1}{J}\mathbf{1}\mathbf{1}^{'},$$

$$\begin{aligned}\mathbb{V}_{\mathbf{p}_i,\mathbb{I}_J,\mathbb{I}_J}\mathbb{V}_i^{-}\mathbb{V}_{\mathbf{p}_i,\mathbb{I}_J,\mathbb{I}_J} &= \left(\mathbb{P}_i - \mathbf{p}_i\mathbf{p}_i{}'\right)\left(\mathbb{P}_i - \mathbf{p}_i\mathbf{p}_i{}'\right)\left[\left(\mathbb{I}_J - \frac{1}{J}\mathbf{1}\mathbf{1}'\right)\mathbb{P}_i^{-1}\left(\mathbb{I}_J - \frac{1}{J}\mathbf{1}\mathbf{1}'\right)\mathbb{P}_i^{-1}\right] \\ &\quad \times \left(\mathbb{P}_i - \mathbf{p}_i\mathbf{p}_i{}'\right)\left(\mathbb{P}_i - \mathbf{p}_i\mathbf{p}_i{}'\right) \\ &= \left(\mathbb{P}_i - \mathbf{p}_i\mathbf{p}_i{}'\right)\left(\mathbb{P}_i - \mathbf{p}_i\mathbf{p}_i{}'\right)\mathbb{P}_i^{-1}\left(\mathbb{I}_J - \frac{1}{J}\mathbf{1}\mathbf{1}'\right)\left(\mathbb{P}_i - \mathbf{p}_i\mathbf{p}_i'\right) \\ &= \left(\mathbb{P}_i - \mathbf{p}_i\mathbf{p}_i{}'\right)\left(\mathbb{P}_i - \mathbf{p}_i\mathbf{p}_i{}'\right)\end{aligned}$$

proving the required result. □

So in this section, the working variance-covariance structure to be used with compositional data has been identified by means of the first order Taylor Series approximation to $\text{Var}(\mathbf{Y}_i)$. By using the invariance property of the estimator $\widehat{\boldsymbol{\gamma}}$, we have also seen that the matrix $\mathbb{V}_{\mathbf{p}_i,\mathbb{I}_J,\mathbb{I}_J}$ may actually be used instead of $\mathbb{V}_{\mathbf{p}_i,\boldsymbol{\Omega},\mathbb{W}}$ to obtain estimates of $\boldsymbol{\gamma}$ where $\boldsymbol{\gamma} = \left(\boldsymbol{\gamma}_1', \ldots, \boldsymbol{\gamma}_{J-1}'\right)'$. The estimates of $\boldsymbol{\gamma}$ arising out of the generalized Wedderburn system of estimating equations will inherit all the desirable properties of estimates obtained using the Liang and Zeger (1986) generalized estimating equations.

2.8 Estimating Standard Errors of $\widehat{\boldsymbol{\gamma}}$

Having identified a working covariance structure that reflects the properties of compositional response variables and which may be used in the generalized Wedderburn system of estimating equations, the next step is that of developing a method to estimate standard errors for $\boldsymbol{\gamma}$. Two variance-covariance estimators will be presented:

- The first estimator is a model-based estimator. Since the same γ estimates are obtained irrespective of the working variance-covariance structure used, a whole family of model-based estimators may actually be defined. However, in analogy to the multivariate regression case, where the GLS estimator is free of the variance-covariance parameters, but the variance-covariance matrix of the estimator is not (e.g. Mardia et al., 1979, p. 173), it is to be expected that an estimator of $\text{Var}(\widehat{\boldsymbol{\gamma}})$ which accounts for the variance-covariance structure of $\mathbf{Y}_i$ leads to better estimated standard errors. The model-based estimator being proposed here is thus based on an estimator of the working variance-covariance structure $\hat{\phi}\widehat{\mathbb{V}_{\mathbf{p}_i,\boldsymbol{\Omega},\mathbb{W}}}$.
- The second estimator is the robust Huber-White sandwich estimator proposed by Liang and Zeger (1986), which allows for potential misspecification in the assumed form $\phi\mathbb{V}_{\mathbf{p}_i,\boldsymbol{\Omega},\mathbb{W}}$.

Pan (2001b) criticizes the estimator proposed by Liang and Zeger (1986) due to the fact that it uses residuals from one case to estimate the true variance-covariance matrix of $\mathbf{Y}_i$. Pan (2001b) proceeds to propose an alternative estimator of the true variance-covariance

matrix. Our model-based estimator will be shown to be in the same form as the estimator in Pan (2001b) and will inherit the same desirable properties of Pan (2001b)'s estimator.

Before proceeding to give detail on the estimators just mentioned, recall that the generalized Wedderburn system of estimating equations that will be used to estimate the parameters $\boldsymbol{\gamma}$ takes the form

$$\sum_{i=1}^{n} \mathbb{D}_i' \mathbb{V}_i^- \left(\mathbf{Y}_i - \mathbf{p}_i\right) = \mathbf{0} \tag{2.64}$$

where $\mathbb{D}_i$ is the matrix of derivatives of $\mathbf{p}_i$ with respect to the parameters $\boldsymbol{\gamma}$, $\mathbb{V}_i$ is the implied variance-covariance structure of the compositional variables, $\mathbb{V}_i^-$ is a generalized inverse of $\mathbb{V}_i$ and $\mathbf{p}_i$ is the mean vector (vector of proportions) corresponding to the response vector $\mathbf{Y}_i$. A generalized inverse is being invoked for $\mathbb{V}_i$, due to this matrix being singular.

Under estimating equations (2.64) and under the assumption that the implied variance-covariance matrix $\phi\mathbb{V}_i$ is equal to the true variance-covariance matrix $\text{Var}\left(\mathbf{Y}_i\right)$, a model-based variance-covariance estimator of $\widehat{\boldsymbol{\gamma}}$ is given by

$$\widehat{\text{Var}\left(\hat{\boldsymbol{\gamma}}\right)}_M = \hat{\phi}\left(\sum_{i=1}^{n} \widehat{\mathbb{D}}_i' \widehat{\mathbb{V}}_i^- \widehat{\mathbb{D}}_i\right)^{-1}, \tag{2.65}$$

where $\widehat{\mathbb{D}}_i$ and $\widehat{\mathbb{V}}_i$ are respectively the matrices $\mathbb{D}_i$ and $\mathbb{V}_i$ estimated using $\widehat{\boldsymbol{\gamma}}$ and $\hat{\phi}$ is the estimated dispersion parameter.

Under estimating equations (2.64), in line with Liang and Zeger (1986), a Huber-White sandwich estimator of $\text{Var}\left(\widehat{\boldsymbol{\gamma}}\right)$ is given by

$$\widehat{\text{Var}\left(\widehat{\boldsymbol{\gamma}}\right)}_R = \left(\sum_{i=1}^{n} \widehat{\mathbb{D}}_i' \widehat{\mathbb{V}}_i^- \widehat{\mathbb{D}}_i\right)^{-1} \left(\sum_{i=1}^{n} \widehat{\mathbb{D}}_i' \widehat{\mathbb{V}}_i^- \widehat{\text{Var}\left(\mathbf{Y}_i\right)} \widehat{\mathbb{V}}_i^- \widehat{\mathbb{D}}_i\right) \left(\sum_{i=1}^{n} \widehat{\mathbb{D}}_i' \widehat{\mathbb{V}}_i^- \widehat{\mathbb{D}}_i\right)^{-1} \tag{2.66}$$

where $\widehat{\mathbb{D}}_i$ and $\widehat{\mathbb{V}}_i$ are as defined for equation (2.65) and $\widehat{\text{Var}\left(\mathbf{Y}_i\right)}$ denotes an estimator of the true typically unknown variance-covariance matrix $\text{Var}\left(\mathbf{Y}_i\right)$. Details on estimating $\text{Var}\left(\mathbf{Y}_i\right)$ will be provided shortly.

The Huber-White estimator, as the name suggests, is attributed to Huber (1967) and White (1980). The papers by Huber (1967) and White (1982), focus on properties of maximum likelihood estimators under model misspecification and White (1980) presents a variance estimator for regression parameters in the presence of heteroscedasticity. It is through the seminal work of Liang and Zeger (1986) that this estimator has been highly popularized in the GEE framework.

2.8.1 Model-Based Estimator in terms of $\widehat{\mathbb{V}_{\mathbf{p}_i,\mathbf{\Omega},\mathbb{W}}}$

Under the assumption that the variance-covariance matrix of $\mathbf{Y}_i$ is given by $\phi\mathbb{V}_{\mathbf{p}_i,\mathbf{\Omega},\mathbb{W}}$, the model-based estimator is given by

$$\widehat{\text{Var}\,(\widehat{\boldsymbol{\gamma}})}_M = \hat{\phi}\left(\sum_{i=1}^{n}\widehat{\mathbb{D}}_i'\left[\widehat{\mathbb{V}_{\mathbf{p}_i,\mathbf{\Omega},\mathbb{W}}}\right]^{-}\widehat{\mathbb{D}}_i\right)^{-1}, \tag{2.67}$$

where $\widehat{\mathbb{D}}_i$ and $\widehat{\mathbb{V}_{\mathbf{p}_i,\mathbf{\Omega},\mathbb{W}}}$ are evaluated at the estimates obtained using the sampled data.

Now, so as to be able to estimate the standard errors of $\widehat{\boldsymbol{\gamma}}$ using estimator (2.67), an estimator of $\phi\mathbb{V}_{\mathbf{p}_i,\mathbf{\Omega},\mathbb{W}}$ is required. Let $\widehat{\mathbf{p}}_i$ denote the estimator of $\mathbf{p}_i$. Given a correct mean model specification, Liang and Zeger (1986) propose to replace $\text{Var}\,(\mathbf{Y}_i)$ in the Huber-White sandwich estimator (2.66) with $(\mathbf{Y}_i-\widehat{\mathbf{p}}_i)(\mathbf{Y}_i-\widehat{\mathbf{p}}_i)'$. Amongst other issues (more detail will be given in Section 2.8.3), the specification of a specific structure for the dependence of the variance-covariance structure on the mean vector is being avoided completely in Liang and Zeger's proposed estimator of $\text{Var}\,(\mathbf{Y}_i)$. An alternative estimator of $\phi\mathbb{V}_{\mathbf{p}_i,\mathbf{\Omega},\mathbb{W}}$ which takes the dependence structure just mentioned into account and which may thus lead to greater efficiency, is developed.

2.8.2 The Development of an Estimator of $\phi\mathbb{V}_{\mathbf{p}_i,\mathbf{\Omega},\mathbb{W}}$

Recall from (2.61) that

$$\phi\mathbb{V}_{\mathbf{p}_i,\mathbf{\Omega},\mathbb{W}} = \left(\mathbb{P}_i-\mathbf{p}_i\mathbf{p}_i'\right)\mathbf{\Sigma}\left(\mathbb{P}_i-\mathbf{p}_i\mathbf{p}_i'\right),$$

which may be rewritten as

$$\phi\mathbb{V}_{\mathbf{p}_i,\mathbf{\Omega},\mathbb{W}} = \left(\mathbb{P}_i-\mathbf{p}_i\mathbf{p}_i'\right)\left(\mathbb{I}_J-\frac{1}{J}\mathbf{1}\mathbf{1}'\right)\mathbf{\Sigma}\left(\mathbb{I}_J-\frac{1}{J}\mathbf{1}\mathbf{1}'\right)\left(\mathbb{P}_i-\mathbf{p}_i\mathbf{p}_i'\right).$$

Let $\left(\mathbb{I}_J-\frac{1}{J}\mathbf{1}\mathbf{1}'\right)\mathbf{\Sigma}\left(\mathbb{I}_J-\frac{1}{J}\mathbf{1}\mathbf{1}'\right)$ be denoted by $\mathbf{\Sigma}^*$ so that

$$\phi\mathbb{V}_{\mathbf{p}_i,\mathbf{\Omega},\mathbb{W}} = \left(\mathbb{P}_i-\mathbf{p}_i\mathbf{p}_i'\right)\mathbf{\Sigma}^*\left(\mathbb{P}_i-\mathbf{p}_i\mathbf{p}_i'\right), \tag{2.68}$$

where $\mathbf{\Sigma}^*$ is free of the index i since it is assumed to be the same for all i.

Estimates of $\mathbb{P}_i$ and $\mathbf{p}_i$ in (2.68) are easily worked out once estimates of $\boldsymbol{\gamma}$ are obtained through the iterative procedure. An estimator for $\mathbf{\Sigma}^*$ needs to be devised.

Since $\mathbf{\Sigma}^*$ is the same for all i, it is natural to develop an estimator for $\mathbf{\Sigma}^*$ which averages

over the information obtained from all i. Let $\widehat{\mathbf{\Sigma}}^*$ be the estimator of $\mathbf{\Sigma}^*$, defined as

$$\widehat{\mathbf{\Sigma}}^* = \frac{1}{n}\sum_{i=1}^{n}\widehat{\mathbf{\Sigma}}_i^*, \tag{2.69}$$

where $\widehat{\mathbf{\Sigma}}_i^*$ is given by

$$\widehat{\mathbf{\Sigma}}_i^* = \left(\mathbb{I}_J - \frac{1}{J}\mathbf{1}\mathbf{1}'\right)\widehat{\mathbb{P}}_i^{-1}\left(\mathbf{Y}_i - \widehat{\mathbf{p}}_i\right)\left(\mathbf{Y}_i - \widehat{\mathbf{p}}_i\right)'\widehat{\mathbb{P}}_i^{-1}\left(\mathbb{I}_J - \frac{1}{J}\mathbf{1}\mathbf{1}'\right). \tag{2.70}$$

The motivation for the expression of $\widehat{\mathbf{\Sigma}}_i^*$ follows.

If $\left(\mathbb{P}_i - \mathbf{p}_i\mathbf{p}_i'\right)^+$ is the Moore-Penrose pseudoinverse of $\mathbb{P}_i - \mathbf{p}_i\mathbf{p}_i'$, defined in (2.52), it may be noted that

$$\begin{aligned}\mathbf{\Sigma}^* &= \left(\mathbb{P}_i - \mathbf{p}_i\mathbf{p}_i'\right)^+\left(\mathbb{P}_i - \mathbf{p}_i\mathbf{p}_i'\right)\mathbf{\Sigma}^*\left(\mathbb{P}_i - \mathbf{p}_i\mathbf{p}_i'\right)\left(\mathbb{P}_i - \mathbf{p}_i\mathbf{p}_i'\right)^+ \\ &= \left(\mathbb{P}_i - \mathbf{p}_i\mathbf{p}_i'\right)^+\phi\mathbb{V}_{\mathbf{p}_i,\mathbf{\Omega},\mathbb{W}}\left(\mathbb{P}_i - \mathbf{p}_i\mathbf{p}_i'\right)^+\end{aligned} \tag{2.71}$$

since $\left(\mathbb{P}_i - \mathbf{p}_i\mathbf{p}_i'\right)^+\left(\mathbb{P}_i - \mathbf{p}_i\mathbf{p}_i'\right) = \mathbb{I}_J - \frac{1}{J}\mathbf{1}\mathbf{1}'$ and $\mathbb{I}_J - \frac{1}{J}\mathbf{1}\mathbf{1}'$ is idempotent.

Consequently, under the assumption that $\phi\mathbb{V}_{\mathbf{p}_i,\mathbf{\Omega},\mathbb{W}} = \mathrm{Var}\left(\mathbf{Y}_i\right)$ (the true variance-covariance matrix of $\mathbf{Y}_i$) and letting $\phi\mathbb{V}_{\mathbf{p}_i,\mathbf{\Omega},\mathbb{W}}$ in (2.71) be estimated by $\left(\mathbf{Y}_i - \widehat{\mathbf{p}}_i\right)\left(\mathbf{Y}_i - \widehat{\mathbf{p}}_i\right)'$, equation (2.71) leads to the estimator

$$\begin{aligned}\widehat{\mathbf{\Sigma}}_i^* &= \left(\widehat{\mathbb{P}}_i - \widehat{\mathbf{p}}_i\widehat{\mathbf{p}}_i'\right)^+\left(\mathbf{Y}_i - \widehat{\mathbf{p}}_i\right)\left(\mathbf{Y}_i - \widehat{\mathbf{p}}_i\right)'\left(\widehat{\mathbb{P}}_i - \widehat{\mathbf{p}}_i\widehat{\mathbf{p}}_i'\right)^+ \\ &= \left(\mathbb{I}_J - \frac{1}{J}\mathbf{1}\mathbf{1}'\right)\widehat{\mathbb{P}}_i^{-1}\left(\mathbf{Y}_i - \widehat{\mathbf{p}}_i\right)\left(\mathbf{Y}_i - \widehat{\mathbf{p}}_i\right)'\widehat{\mathbb{P}}_i^{-1}\left(\mathbb{I}_J - \frac{1}{J}\mathbf{1}\mathbf{1}'\right).\end{aligned} \tag{2.72}$$

Now, in the generalized linear modeling framework, a Pearson residual is defined by $\frac{Y_{ij}-\hat{p}_{ij}}{\sqrt{V(\hat{p}_{ij})}}$ where $\hat{p}_{ij}$ is the fitted value of Y_{ij} and $V(\cdot)$ is a variance function. The general idea behind a Pearson residual is that of converting the difference $\mathbf{Y}_i - \mathbf{p}_i$ into a vector of standardized residuals whose mean vector and variance-covariance matrix do not depend on $\mathbf{p}_i$. From the mean-model specification assumed under the generalized Wedderburn approach, we have that $E\left(\mathbf{Y}_i\right) = \mathbf{p}_i$ so that $E\left(\mathbf{Y}_i - \mathbf{p}_i\right) = \mathbf{0}$. From Section 2.7, the working variance-covariance structure of $\mathbf{Y}_i$ is given by $\left(\mathbb{P}_i - \mathbf{p}_i\mathbf{p}_i'\right)\mathbf{\Sigma}\left(\mathbb{P}_i - \mathbf{p}_i\mathbf{p}_i'\right)$ and

$$\begin{aligned}\mathrm{Var}\left(\left(\mathbb{I}_J - \frac{1}{J}\mathbf{1}\mathbf{1}'\right)\mathbb{P}_i^{-1}\left(\mathbf{Y}_i - \mathbf{p}_i\right)\right) &= \left(\mathbb{I}_J - \frac{1}{J}\mathbf{1}\mathbf{1}'\right)\mathbb{P}_i^{-1}\left(\mathbb{P}_i - \mathbf{p}_i\mathbf{p}_i'\right)\mathbf{\Sigma}\left(\mathbb{P}_i - \mathbf{p}_i\mathbf{p}_i'\right)\mathbb{P}_i^{-1} \\ &\quad\times\left(\mathbb{I}_J - \frac{1}{J}\mathbf{1}\mathbf{1}'\right) \\ &= \left(\mathbb{I}_J - \frac{1}{J}\mathbf{1}\mathbf{1}'\right)\mathbf{\Sigma}\left(\mathbb{I}_J - \frac{1}{J}\mathbf{1}\mathbf{1}'\right),\end{aligned}$$

since $\left(\mathbb{I}_J - \frac{1}{J}\mathbf{1}\mathbf{1}'\right)\mathbb{P}_i^{-1}\left(\mathbb{P}_i - \mathbf{p}_i\mathbf{p}_i'\right) = \mathbb{I}_J - \frac{1}{J}\mathbf{1}\mathbf{1}'$.

Thus, by premultiplyling $\mathbf{Y}_i - \mathbf{p}_i$ by $\left(\mathbb{I}_J - \frac{1}{J}\mathbf{1}\mathbf{1}'\right)\mathbb{P}_i^{-1}$, we have achieved a random vector whose mean vector and variance-covariance matrix are not dependent on $\mathbf{p}_i$, showing that under the generalized Wedderburn approach, the Pearson residual vector should take the form

$$\left(\mathbb{I}_J - \frac{1}{J}\mathbf{1}\mathbf{1}'\right)\widehat{\mathbb{P}}_i^{-1}\left(\mathbf{Y}_i - \widehat{\mathbf{p}}_i\right). \tag{2.73}$$

On comparing (2.70) with expression (2.73), it may thus be noted that the estimator $\widehat{\mathbf{\Sigma}}_i^*$ is a 'squared Pearson residual' matrix for case i.

Having obtained $\widehat{\mathbf{\Sigma}}^*$, we then estimate $\phi\mathbb{V}_{\mathbf{p}_i,\mathbf{\Omega},\mathbb{W}}$ by

$$\widehat{\phi\mathbb{V}_{\mathbf{p}_i,\mathbf{\Omega},\mathbb{W}}} = \left(\widehat{\mathbb{P}}_i - \widehat{\mathbf{p}}_i\widehat{\mathbf{p}}_i'\right)\frac{1}{n}\sum_{i=1}^{n}\widehat{\mathbf{\Sigma}}_i^*\left(\widehat{\mathbb{P}}_i - \widehat{\mathbf{p}}_i\widehat{\mathbf{p}}_i'\right). \tag{2.74}$$

The denominator in $\widehat{\mathbf{\Sigma}}^* = \frac{1}{n}\sum_{i=1}^{n}\widehat{\mathbf{\Sigma}}_i^*$ may also be adjusted from n to $(n-(p+1))$ to cater for the γ coefficients being estimated through the generalized Wedderburn estimating equations (2.64). In what follows, the estimator $\widehat{\phi\mathbb{V}_{\mathbf{p}_i,\mathbf{\Omega},\mathbb{W}}}$ will always be considered in terms of the adjusted denominator.

Focusing on the case $J = 2$

To get a better insight into the role that (2.74) plays in the variance estimator (2.67) for the generalized Wedderburn system, consider the case $J = 2$. Due to the sum-constraint,

$$\widehat{\mathbf{\Sigma}}_i^* = \frac{1}{4}\frac{(Y_{i1} - \hat{p}_{i1})^2}{\hat{p}_{i1}^2(1-\hat{p}_{i1})^2}\begin{pmatrix} 1 & -1 \\ -1 & 1 \end{pmatrix}$$

so that

$$\widehat{\mathbf{\Sigma}}^* = \frac{1}{4(n-(p+1))}\left[\sum_{i=1}^{n}\frac{(Y_{i1} - \hat{p}_{i1})^2}{\hat{p}_{i1}^2(1-\hat{p}_{i1})^2}\right]\begin{pmatrix} 1 & -1 \\ -1 & 1 \end{pmatrix}$$

leading to

$$\widehat{\mathrm{Var}(\widehat{\gamma})}_M = \left[\frac{1}{(n-(p+1))}\sum_{i=1}^{n}\frac{(Y_{i1} - \hat{p}_{i1})^2}{\hat{p}_{i1}^2(1-\hat{p}_{i1})^2}\right]\left(\mathbb{X}'\mathbb{X}\right)^{-1}. \tag{2.75}$$

For details on the derivation of (2.75), refer to Appendix A.

Now consider the model used by Wedderburn (1974) for barley leaf data, in which a logit link function has been used together with a mean-variance relationship defined by

$V(p_{i1}) = p_{i1}^2(1-p_{i1})^2$. Let $\widehat{\text{Var}(\widehat{\boldsymbol{\gamma}})}_W$ denote the estimated variance under Wedderburn's model. The estimated variance $\widehat{\text{Var}(\widehat{\boldsymbol{\gamma}})}_W$ is given by

$$\widehat{\text{Var}(\widehat{\boldsymbol{\gamma}})}_W = \widehat{\phi}_W \left(\mathbb{X}'\mathbb{X}\right)^{-1} \tag{2.76}$$

where $\widehat{\phi}_W$ is the estimator of the dispersion parameter ϕ_W.

On comparing (2.76) with (2.75), it may be noticed that the two estimators take the same form, with estimator (2.75) providing the explicit expression, $\frac{1}{n-(p+1)} \sum_{i=1}^{n} \frac{(Y_{i1}-\hat{p}_{i1})^2}{\hat{p}_{i1}^2(1-\hat{p}_{i1})^2}$, to estimate the dispersion parameter ϕ_W. This estimator for ϕ_W is none other than the method of moments estimator used to estimate dispersion parameters in quasi-likelihood estimation. So, for $J = 2$,

- in Section 2.5 it was shown that the generalized Wedderburn system produces γ estimates that are equivalent to those obtained from Wedderburn's estimating equations and
- from this section, it has been shown that the general method devised to estimate standard errors under the generalized Wedderburn system, also agrees with that of Wedderburn (1974).

It is also important to note that the newly developed estimator $\widehat{\boldsymbol{\Sigma}}^*$ is an estimator which 'borrows strength across subjects' (Liang and Zeger, 1986). Liang and Zeger (1986) apply such methodology in their proposed estimator for an unstructured correlation matrix (Liang and Zeger, 1986, eq. (9)) but fail to use the same idea in the specification of an estimator for the true variance-covariance matrix.

2.8.3 The Liang and Zeger (1986) Robust Sandwich Estimator

As mentioned in the previous section, given a correct mean model specification, Liang and Zeger (1986) propose to estimate $\text{Var}(\widehat{\boldsymbol{\gamma}})$ by using the Huber-White sandwich estimator (2.66) where $\text{Var}(\mathbf{Y}_i)$ is estimated by $(\mathbf{Y}_i - \widehat{\mathbf{p}}_i)(\mathbf{Y}_i - \widehat{\mathbf{p}}_i)'$. This approach leads to a robust estimator which gives consistent estimates of the standard errors (Liang and Zeger, 1986). Several researchers have however raised issues on the use of this sandwich estimator, particularly for use with small sample sizes (Emrich and Piedmonte, 1992; Drum and McCullagh, 1993; Mancl and DeRouen, 2001; Pan, 2001b; Gosho et al., 2014).

The difference between observed and fitted values tends to be smaller than the difference between the true and fitted values (Mancl and DeRouen, 2001). So since the estimate of $\text{Var}(\mathbf{Y}_i)$ is defined using the residuals of only one case, it is to be expected that for small sample sizes, the sandwich estimator underestimates $\widehat{\text{Var}(\widehat{\boldsymbol{\gamma}})}$. A downward bias in the variance estimate leads to inflated Type I errors (Emrich and Piedmonte, 1992) and elevated values of the robust Wald test, leading to lowered coverage probabilities of the

corresponding confidence intervals. The robust Wald test is used for testing significance of the model coefficients. More detail on this test will be given in Section 2.9.1.2. The lack of accuracy used in estimating $\text{Var}(\mathbf{Y}_i)$, also leads to an increase in variability of the sandwich estimator. Kauermann and Carroll (2001) show a decrease in efficiency of the sandwich estimator in comparison to a model-based estimator with Poisson and Binomial data with known dispersion parameter. In reference to using the sandwich estimator with small sample sizes, Drum and McCullagh (1993) state that 'by failing to pool variances [pooling information from all cases being considered in the study], all risk of contamination is avoided. But the cost of protection seems high'. Drum and McCullagh (1993) go on to state their preference for a model-based estimator of the variance, 'unless there is enough reason to believe that the assumed variance function is substantially incorrect'.

There is disagreement about the use of $(\mathbf{Y}_i - \widehat{\mathbf{p}}_i)(\mathbf{Y}_i - \widehat{\mathbf{p}}_i)'$ as an estimator of $\text{Var}(\mathbf{Y}_i)$ is argued upon by other researchers (Pan, 2001b; Gosho et al., 2014). Due to the fact that the estimator is based on residuals from one subject/object, Pan (2001b) and Gosho et al. (2014) state that $(\mathbf{Y}_i - \widehat{\mathbf{p}}_i)(\mathbf{Y}_i - \widehat{\mathbf{p}}_i)'$ is in fact neither consistent nor efficient.

Different approaches have been undertaken in order to tackle the problems associated with Liang and Zeger's (1986) sandwich estimator. One approach takes into account the variability of the estimator. Fay and Graubard (2001) and Pan and Wall (2002) constructed Wald-type tests based on the F and t distribution (instead of being based on the chi-square or normal distribution) for testing single or multiple parameters. Another approach takes into account of the bias of the sandwich estimator arising due to the residuals. The aim of the latter approach is that of reducing the bias by developing a new estimator of $\text{Var}(\mathbf{Y}_i)$. Mancl and DeRouen (2001) suggest an estimator which is based on the idea of reducing the bias of the residual estimator $(\mathbf{Y}_i - \widehat{\mathbf{p}}_i)(\mathbf{Y}_i - \widehat{\mathbf{p}}_i)'$. Pan (2001b) proposes to estimate $\text{Var}(\mathbf{Y}_i)$ by exploiting the implied structure in the working variance-covariance matrix.

More detail on the estimator in Pan (2001b) will be given in the following section. Special attention will be given to this estimator since with the specification of two additional assumptions, Pan's estimator will be shown to be more efficient than Liang and Zeger's estimator. It will also be shown that the variance estimator (2.74), proposed for use with compositional data, is in the form of Pan's estimator and the properties of Pan's estimator are also inherited by estimator (2.74).

2.8.4 Estimating $\text{Var}(\mathbf{Y}_i)$ as per Pan (2001b)

In Liang and Zeger (1986), the working variance-covariance structure is assumed to be of the form $\phi\mathbb{V}_i = \phi\mathbb{A}_i^{\frac{1}{2}}\mathbb{W}(\boldsymbol{\alpha})\mathbb{A}_i^{\frac{1}{2}}$, where an unstructured working correlation matrix $\mathbb{W}(\boldsymbol{\alpha})$

is estimated using

$$\widehat{\mathbb{W}}(\boldsymbol{\alpha}) = \frac{1}{n\hat{\phi}} \sum_{i=1}^{n} \widehat{\mathbb{A}}_i^{-\frac{1}{2}} (\mathbf{Y}_i - \widehat{\mathbf{p}}_i) (\mathbf{Y}_i - \widehat{\mathbf{p}}_i)' \widehat{\mathbb{A}}_i^{-\frac{1}{2}}. \tag{2.77}$$

Under the assumptions that the true variance-covariance structure needs to be modeled correctly and that there is a common correlation structure $\mathbb{W}(\boldsymbol{\alpha})$ amongst all cases, Pan (2001b) proposes to estimate $\text{Var}(\mathbf{Y}_i)$ using

$$\widehat{\text{Var}(\mathbf{Y}_i)} = \hat{\phi} \widehat{\mathbb{A}}_i^{\frac{1}{2}} \widehat{\mathbb{W}}(\boldsymbol{\alpha}) \widehat{\mathbb{A}}_i^{\frac{1}{2}} \tag{2.78}$$

where $\widehat{\mathbb{W}}(\boldsymbol{\alpha})$ is given by estimator (2.77) leading to

$$\widehat{\text{Var}(\mathbf{Y}_i)} = \widehat{\mathbb{A}}^{\frac{1}{2}} \left[\frac{1}{n} \sum_{i=1}^{n} \widehat{\mathbb{A}}_i^{-\frac{1}{2}} (\mathbf{Y}_i - \widehat{\mathbf{p}}_i) (\mathbf{Y}_i - \widehat{\mathbf{p}}_i)' \widehat{\mathbb{A}}_i^{-\frac{1}{2}} \right] \widehat{\mathbb{A}}^{\frac{1}{2}}. \tag{2.79}$$

Analogous to the variance estimator (2.74), that has been proposed for use in modeling compositional data, the estimator proposed by Pan (2001b) also 'borrows strength across subjects' (Liang and Zeger, 1986) to estimate $\text{Var}(\mathbf{Y}_i)$. Estimator (2.74) is in fact expressed in the same form as estimator (2.79), with the former estimator having the matrix $\mathbb{P}_i - \mathbf{p}_i \mathbf{p}_i'$ instead of $\mathbb{A}_i^{\frac{1}{2}}$.

Since Pan's estimator is obtained by pooling information from the different cases, Pan (2001b) states that it is to be expected that the estimator proposed by Liang and Zeger (1986) has lower efficiency and goes on to prove an asymptotic result showing better efficiency for estimator (2.79). This result is presented next.

To be able to present Pan (2001b)'s result, focus is directed towards the middle part of the sandwich estimator, $\sum_{i=1}^{n} \left(\widehat{\mathbb{D}}_i' \widehat{\mathbb{V}}_i^{-} \widehat{\text{Var}(\mathbf{Y}_i)} \widehat{\mathbb{V}}_i^{-} \widehat{\mathbb{D}}_i \right)$, under both Liang and Zeger (1986)'s estimator and Pan (2001b)'s estimator. Let the two respective middle parts be denoted by M_{LZ} and M_P respectively. Treating matrices $\widehat{\mathbb{D}}_i$, $\widehat{\mathbb{V}}_i$ and $\widehat{\mathbb{A}}_i$ as fixed, Pan (2001b) shows that asymptotically, under mild regularity conditions,

$$\text{Var}(\text{vec}(M_{LZ})) \geq \text{Var}(\text{vec}(M_P)) \tag{2.80}$$

where for some matrix $\mathbb{M}$, $\text{vec}(\mathbb{M})$ stacks all the columns of $\mathbb{M}$ into one vector.

Let Liang and Zeger (1986) and Pan (2001b) sandwich estimators be denoted by $\widehat{\text{Var}(\widehat{\boldsymbol{\gamma}})}_{LZ}$ and $\widehat{\text{Var}(\widehat{\boldsymbol{\gamma}})}_P$ respectively. Result (2.80) leads to

$$\text{Var}\left(\text{vec}\left(\widehat{\text{Var}(\widehat{\boldsymbol{\gamma}})}_{LZ}\right)\right) \geq \text{Var}\left(\text{vec}\left(\widehat{\text{Var}(\widehat{\boldsymbol{\gamma}})}_P\right)\right) \tag{2.81}$$

showing that by sacrificing some of the robustness provided by the estimator in Liang and

Zeger (1986), the estimator suggested by Pan (2001b) achieves higher efficiency.

By substituting the matrix $\mathbb{A}_i^{\frac{1}{2}}$, in the proof for the above result presented in Pan (2001b, p. 905), by $\mathbb{P}_i - \mathbf{p}_i\mathbf{p}_i{}'$, it may be shown that the result also holds for estimator (2.74). This shows that asymptotically, the proposed estimator (2.74) estimates the standard errors of $\widehat{\boldsymbol{\gamma}}$ more efficiently than Liang and Zeger's estimator.

Although the result has been proved for n tending to infinity, the simulation study carried out by Pan (2001b) using sample sizes of 10 and 40, suggests that greater efficiency for Pan (2001b) estimator holds even under small sample sizes. Also, as per Pan (2001b), using (2.79) as an estimator of the true variance-covariance matrix yields a consistent estimator of $\text{Var}\,(\hat{\boldsymbol{\gamma}})$. Gosho et al. (2014) propose to carry out a degrees of freedom correction, so as to achieve an unbiased estimator of $\text{Var}\,(\mathbf{Y}_i)$. This is the same correction that has been mentioned in conjunction with the variance estimator (2.74).

2.9 Testing the Quality of Fit of the Model

Once a model is fit to a data set, the goodness of fit of the model needs to be checked. Techniques for model testing and model selection in the context of modeling independent data, as in generalized linear modeling, are widely available (McCullagh and Nelder, 1989, p. 33-40). One particular measure that is typically used for model comparison of nested models in generalized linear modeling is the difference in deviance. The deviance is defined as $-2\times$ the log-likelihood ratio of the model being tested (reduced model) versus the full model (also known as saturated model, since this model has as many parameters as the size of the sample being considered). Information criteria, such as AIC and BIC are also used for testing of models fitted to non-correlated data. These latter criteria may also be used with non-nested models. Other measures which may be used to test the quality of fit of a generalized linear model include the Wald test, the score test and residuals. The Wald test statistic and the score statistic are used to either test for significance of the model coefficients or to test for significance of sets of contrasts between the model coefficients. On testing for significance of the r^{th} model coefficient, $(r = 1, \ldots, p+1)$, the null hypothesis takes the form $H_0 : \beta_r = 0$. On testing for significance of a set of contrasts, the null hypothesis takes the form $H_0 : \mathbb{L}\boldsymbol{\beta} = \mathbf{0}$ for some $(q \times (p+1))$ matrix $\mathbb{L}$, with $q \leq p$.

In the context of generalized estimating equations, testing for lack of fit becomes more complicated since only the first two moments and a working correlation structure are specified. The just mentioned difference in deviance and the information criteria rely on the full specification of the likelihood function. The score statistic and the Wald statistic are based on the score function, which is again only available under the specification of the likelihood function. Measures used to test quality of fit of a model when GEE is used will be described in Section 2.9.1. Section 2.9.2 will then give detail on those measures that are deemed to be suitable for testing quality of fit of the multivariate logit model for

compositional data.

Note that in this introductory part of this section and also in Section 2.9.1, the notation $\boldsymbol{\beta}$ is used to represent a general set of model coefficients. In Section 2.9.2, the notation will be changed in accordance with the multivariate logit model.

2.9.1 Testing the Quality of Fit of the Model when GEE is Used

Despite the fact that a quasi-score function has many properties that are similar to the score function arising out of a regular likelihood function (more detail on this has been provided in Section 2.4.1), there is a property which the quasi-score function might not satisfy. The score function is by definition the derivative of the log-likelihood function with respect to the model parameters. This also works the other way round. The log-likelihood function can be achieved by integrating the score function with respect to the parameters. However, unless the matrix of derivatives $\frac{\partial \mathbf{U}}{\partial \boldsymbol{\beta}}$ is symmetric, the line integral of the quasi-score function will depend on the path chosen (McCullagh, 1990, p. 284). Only if the line integral is path independent can there be a unique quasi-log likelihood function. Without a unique quasi-log likelihood function,'it is difficult to distinguish good roots from bad roots' (Li and McCullagh, 1994), especially in the case where the estimating equations have multiple roots. As per Li and McCullagh (1994), 'a quasi-score function frequently fails to have a symmetric derivative matrix.' McLeish and Small (1992) developed a method in which the likelihood ratio is projected onto the linear space spanned by the product of independent observations, $\{y_1 - m_1(\boldsymbol{\beta}), \ldots, y_n - m_n(\boldsymbol{\beta})\}$. McLeish and Small (1992) showed that if the quasi log-likelihood function exists, the logarithm of the projection leads to first order equivalent inferences to those obtained from the quasi log-likelihood function. Li (1993) also makes use of projections, and obtains a deviance function which approximates the quasi log-likelihood ratio. Li and McCullagh (1994) proposed a method of projecting the quasi-score function onto a class of estimating functions that satisfy the symmetric property in the derivatives $\frac{\partial \mathbf{U}}{\partial \boldsymbol{\beta}}$, yielding a unique integral for the projected quasi-score. Hanfelt and Liang (1995) then developed an approximate likelihood ratio for regular unbiased estimating equations (generalized estimating equations fall into this class) using two different methods; the first based on a quasi-likelihood approach and the second based on projections. The two methods use only information that is required to set up the estimating equations. The quasi-likelihood approach involves a line-integral over the parameter space of the estimating equations. The projection approach works by projecting the log-likelihood ratio onto the space spanned by a linear combination of estimating equations. If the estimating equations used are quasi-score equations, this projection method reduces to the projection likelihood ratio devised by Li (1993) for use with quasi-likelihood estimation. Hanfelt and Liang (1995) have also shown that asymptotically correct hypothesis tests are achieved with either one of the two approaches used to approximate the likelihood ratio, irrespective of the path of integration chosen for the quasi-likelihood approach and intermediate points needed to work out the projection based approach. Despite being

promising, to my knowledge, this area of research has not been developed enough to produce tools that could be used for testing quality of fit of a model whose coefficients have been estimated using generalized estimating equations/quasi-likelihood estimation.

The statistic that appears to be the most widely used for model selection when using generalized estimating equations is the QIC (Pan, 2001a).

2.9.1.1 The Quasi Information Criterion (QIC)

The QIC is an information criterion, analogous to the AIC, but which is modified so that it makes use of i) the quasi-likelihood function under the working independence model, ii) the estimates of the model coefficients under the chosen working correlation matrix and iii) the robust variance estimator proposed by Liang and Zeger (1986).

Under the working assumption of independence for all cases i and for all response variables j, Pan (2001a) first presents the log quasi-likelihood function $Q\left(\boldsymbol{\beta}, \mathbb{I}_J, \mathbf{y}_1, \ldots, \mathbf{y}_n\right)$:

$$Q\left(\boldsymbol{\beta}, \mathbb{I}_J, \mathbf{y}_1, \ldots, \mathbf{y}_n\right) = \sum_{i=1}^{n} \sum_{j=1}^{J} Q\left(\boldsymbol{\beta}, \mathbb{I}_J; Y_{ij}\right). \tag{2.82}$$

The QIC is then defined as

$$QIC(\mathbb{W}) = -2Q\left(\widehat{\boldsymbol{\beta}}\left(\mathbb{W}\right); \mathbb{I}_J, \mathbf{y}_1, \ldots, \mathbf{y}_n\right) - 2trace\left(\left[\widehat{\text{Var}\left(\widehat{\boldsymbol{\beta}}\right)}_M\right]^{-1} \widehat{\text{Var}\left(\widehat{\boldsymbol{\beta}}\right)}_R\right), \tag{2.83}$$

where

- $Q\left(\widehat{\boldsymbol{\beta}}\left(\mathbb{W}\right); \mathbb{I}_J, \mathbf{y}_1, \ldots, \mathbf{y}_n\right)$ is the log quasi-likelihood function (2.82) evaluated at estimates of $\boldsymbol{\beta}$ obtained under the working correlation structure $\mathbb{W}$,
- $\widehat{\text{Var}\left(\widehat{\boldsymbol{\beta}}\right)}_M$ is the model-based variance estimator $\widehat{\phi}\left(\sum_{i=1}^{n} \widehat{\mathbb{D}}_i' \widehat{\mathbb{V}}_i^{-1} \widehat{\mathbb{D}}_i\right)^{-1}$ where $\mathbb{D}_i$ is the matrix of derivatives with elements $\partial m_i\left(\boldsymbol{\beta}\right)/\partial \beta_k$, $(k = 1, \ldots, p+1)$, $\mathbb{V}_i$ is the 'working' variance-covariance matrix, $\widehat{\mathbb{D}}_i$ and $\widehat{\mathbb{V}}_i$ are the estimates of $\mathbb{D}_i$ and $\mathbb{V}_i$ respectively, evaluated using the β estimates obtained under an independence working correlation structure,
- $\widehat{\text{Var}\left(\widehat{\boldsymbol{\beta}}\right)}_R$ is the robust variance estimator proposed by Liang and Zeger (1986).

The QIC may be used for both variable selection and working correlation matrix selection. The model yielding the smallest QIC is deemed to be the best fitting model. Hin and Wang (2009) suggested using a modified version of the QIC, the so-called CIC, which makes use of the second term in the QIC only. Hin and Wang (2009) show that the CIC leads to improvement in the choice of the working correlation structure. Gosho et al. (2011) considered improved versions of QIC and CIC for choosing the best working correlation

structure. The improved versions make use of alternative robust variance estimators than the one suggested by Liang and Zeger (1986). In the analysis carried out by Gosho et al. (2011), it was shown that the modified versions of QIC and CIC based on Pan (2001b) variance estimator had the best overall performance in selecting the best working correlation structure.

Under a GEE framework, it is also possible to test for significance of model coefficients by means of four tests - a generalized version of the Wald test statistic, a 'working' Wald statistic, a generalized version of the score statistic and a 'working' score statistic. These test statistics are presented in Rotnizky and Jewell (1990). Boos (1992) discusses these tests from a general perspective.

2.9.1.2 The Generalized Wald Tests

The null hypothesis being tested under the four tests is the same as for the standard Wald and score test. The simplest hypothesis to test is as follows. Suppose that we wish to test the null hypothesis $H_0 : \boldsymbol{\beta}^{(1)} = \boldsymbol{\beta}^0$ where $\boldsymbol{\beta}^{(1)}$ is an r-vector and subset of the $(p+1)$-vector of model coefficients $\boldsymbol{\beta}$, and $\boldsymbol{\beta}^0$ denotes the r-vector of values assigned to $\boldsymbol{\beta}^{(1)}$ under H_0.

The Generalized Wald Statistic

The generalized Wald statistic is given by

$$T_W = \left(\widehat{\boldsymbol{\beta}}^{(1)} - \boldsymbol{\beta}^0\right)' \left[\widehat{\text{Var}\left(\widehat{\boldsymbol{\beta}}^{(1)}\right)}_R\right]^{-1} \left(\widehat{\boldsymbol{\beta}}^{(1)} - \boldsymbol{\beta}^0\right), \tag{2.84}$$

where $\widehat{\text{Var}\left(\widehat{\boldsymbol{\beta}}^{(1)}\right)}_R$ is the $(r \times r)$ submatrix of Liang and Zeger (1986) robust variance-covariance matrix for $\widehat{\boldsymbol{\beta}}$.

Since as per Liang and Zeger (1986), the GEE estimator $\widehat{\boldsymbol{\beta}}$ is asymptotically multivariate normal, then under H_0, T_W is chi-square distributed with r degrees of freedom.

As per Rotnizky and Jewell (1990), the estimation of $\text{Var}\left(\widehat{\boldsymbol{\beta}}^{(1)}\right)_R$ might be unstable in the presence of a small sample size n with a large number of repeated observations for each case i: 'the test sizes are inflated and the corresponding confidence intervals have lower coverage probabilities' (Guo et al., 2005). Other issues related to using Liang and Zeger's sandwich estimator have been mentioned in Section 2.8.3. The bias-corrected variance estimator proposed by Mancl and DeRouen (2001) and the variance estimator proposed by Pan (2001b) provide alternative means of estimating $\text{Var}\left(\widehat{\boldsymbol{\beta}}^{(1)}\right)$.

The 'Working' Wald Statistic

Under the assumption that the working correlation structure is the true correlation structure, Rotnizky and Jewell (1990) propose to use the modified 'working' Wald statistic T_W^* as an alternative to the generalized version of the Wald test statistic. The statistic T_W^* is defined by

$$T_W^* = \left(\widehat{\boldsymbol{\beta}}^{(1)} - \boldsymbol{\beta}^0\right)' \left[\widehat{\text{Var}\left(\widehat{\boldsymbol{\beta}}^{(1)}\right)}_M\right]^{-1} \left(\widehat{\boldsymbol{\beta}}^{(1)} - \boldsymbol{\beta}^0\right), \tag{2.85}$$

where $\widehat{\text{Var}\left(\widehat{\boldsymbol{\beta}}^{(1)}\right)}_M$ is the $(r \times r)$ submatrix of the model-based variance estimator for $\widehat{\boldsymbol{\beta}}$, $\widehat{\phi}\left(\sum_{i=1}^n \widehat{\mathbb{D}}_i' \widehat{\mathbb{V}}_i^{-1} \widehat{\mathbb{D}}_i\right)^{-1}$. Under mild regularity conditions and provided that the marginal mean model specification is correct, Rotnizky and Jewell (1990, Theorem 1) proved that

$$T_W^* = \sum_{j=1}^r c_j^2 \chi_j^2 + o_p(1), \tag{2.86}$$

where χ_j^2 $(j = 1, \ldots, r)$ are independent chi-square distributed random variables with 1 degree of freedom and $c_1 \geq \ldots \geq c_r$ are the eigenvalues of $\mathbb{Q} = \mathbb{Q}_0^{-1}\mathbb{Q}_1 + o^{r\times r}(1)$ with

$$\mathbb{Q}_0 = \sum_{i=1}^n \tilde{\mathbb{D}}_i' \mathbb{V}_i^{-1} \tilde{\mathbb{D}}_i^{-1},$$

$$\mathbb{Q}_1 = \sum_{i=1}^n \tilde{\mathbb{D}}_i' \mathbb{V}_i^{-1} \text{Var}(\mathbf{Y}_i) \mathbb{V}_i^{-1} \tilde{\mathbb{D}}_i^{-1},$$

where

$$\tilde{\mathbb{D}}_i' = \left\{ \begin{array}{ll} \mathbb{D}_i^{(1)} - \mathbb{D}_i^{(2)} \left(\sum_{i=1}^n \mathbb{D}_i^{(2)'} \mathbb{V}_i^{-1} \mathbb{D}_i^{(2)}\right)^{-1} \left(\sum_{i=1}^n \mathbb{D}_i^{(2)'} \mathbb{V}_i^{-1} \mathbb{D}_i^{(1)}\right), & \text{for } r < p+1 \\ \mathbb{D}_i & \text{for } r = p. \end{array} \right\}$$

Here $\mathbb{D}_i^{(1)}$ is the $J \times r$ matrix making up the first r columns in $\mathbb{D}_i$, $\mathbb{D}_i^{(2)}$ is the $J \times ((p+1) - r)$ matrix making up the last $((p+1) - r)$ columns in $\mathbb{D}_i$ and $o^{r\times r}(1)$ is an $r \times r$ matrix with $o(1)$ elements as n tends to infinity. If a sequence of random variables X_n satisfies $X_n = o_p(1)$, then X_n converges to zero in probability as n tends to infinity. If a sequence of random variables X_n satisfies $X_n = o(1)$, then X_n tends to 0 as n tends to infinity. Note that the last two sentences define standard asymptotic notation.

If the chosen working correlation structure is the true correlation structure, $\mathbb{Q}$ reduces to $\mathbb{I}_J$. So $c_j = 1$ for all j and thus T_W^* becomes asymptotically chi-square distributed with r degrees of freedom.

Now as per Rotnizky and Jewell (1990), a major disadvantage of the Wald test statistic is that it is not invariant to nonlinear transformations of the model coefficients. In such a situation, the generalized score statistic and the 'working' score test statistic are bet-

ter alternatives to the two Wald test statistics presented in this section. The generalized score statistic and the 'working' score test statistic are both invariant to differentiable transformations of the model coefficients (Rotnizky and Jewell, 1990). So focus will now be directed towards the generalized score statistic and the 'working' score test statistic. However, since typically only linear transformations of the model coefficients are required for testing for significance of model coefficients in the multivariate logit model for compositional data, both Wald and score tests may be considered in analyzing compositional data sets.

2.9.1.3 The Generalized Score Tests

Let $\boldsymbol{\beta} = \left(\boldsymbol{\beta}^{(1)'}, \boldsymbol{\beta}^{(2)'}\right)'$ where $\boldsymbol{\beta}^{(1)}$ and $\boldsymbol{\beta}^{(2)}$ are an r-vector and a $((p+1)-r)$-vector of parameters respectively. Also let

$$\mathbf{U}\left(\boldsymbol{\beta}^{(1)}, \boldsymbol{\beta}^{(2)}\right) = \left(\mathbf{U}^{(1)}\left(\boldsymbol{\beta}^{(1)}, \boldsymbol{\beta}^{(2)}\right)', \mathbf{U}^{(2)}\left(\boldsymbol{\beta}^{(1)}, \boldsymbol{\beta}^{(2)}\right)'\right)',$$

where $\mathbf{U}\left(\boldsymbol{\beta}^{(1)}, \boldsymbol{\beta}^{(2)}\right)$ is the $(p+1)$-vector of estimating functions for $\boldsymbol{\beta}$, $\mathbf{U}^{(1)}\left(\boldsymbol{\beta}^{(1)}, \boldsymbol{\beta}^{(2)}\right)$ is made up of the estimating functions corresponding to $\boldsymbol{\beta}^{(1)}$, whilst $\mathbf{U}^{(2)}\left(\boldsymbol{\beta}^{(1)}, \boldsymbol{\beta}^{(2)}\right)$ is made up of the estimating functions corresponding to $\boldsymbol{\beta}^{(2)}$.

The Generalized Score Test Statistic

For the generalized score test statistic, once again suppose that we wish to test the null hypothesis $H_0 : \boldsymbol{\beta}^{(1)} = \boldsymbol{\beta}^0$ where $\boldsymbol{\beta}^0$ denotes the r-vector of values assigned to $\boldsymbol{\beta}^{(1)}$ under H_0. The generalized score test proceeds by solving $\mathbf{U}^{(2)}\left(\boldsymbol{\beta}^{(1)}, \boldsymbol{\beta}^{(2)}\right) = \mathbf{0}$ to get an estimate of $\boldsymbol{\beta}^{(2)}$. Let this estimate be denoted by $\widehat{\boldsymbol{\beta}}^{(2)}$. The test statistic is then given by

$$T_S = \mathbf{U}^{(1)}\left(\boldsymbol{\beta}^0, \widehat{\boldsymbol{\beta}}^{(2)}\right)\widehat{\mathbb{V}}_0^{-1}\mathbf{U}^{(1)}\left(\boldsymbol{\beta}^0, \widehat{\boldsymbol{\beta}}^{(2)}\right), \tag{2.87}$$

where

$$\widehat{\mathbb{V}}_0 = \left(\left[\widehat{\text{Var}\left(\widehat{\boldsymbol{\beta}}^{(1)}\right)}_M^{-1}\right]\left[\widehat{\text{Var}\left(\widehat{\boldsymbol{\beta}}^{(1)}\right)}_R\right]\left[\widehat{\text{Var}\left(\widehat{\boldsymbol{\beta}}^{(1)}\right)}_M^{-1}\right]\right)_{\boldsymbol{\beta}=\left(\boldsymbol{\beta}^0, \widehat{\boldsymbol{\beta}}^{(2)}\right)}$$

in which $\widehat{\text{Var}\left(\widehat{\boldsymbol{\beta}}^{(1)}\right)}_M$ is the model-based estimator defined for T_W^* and $\widehat{\text{Var}\left(\widehat{\boldsymbol{\beta}}^{(1)}\right)}_R$ is Liang and Zeger (1986) robust variance-covariance estimator defined for T_W.

As per Rotnizky and Jewell (1990), under mild regularity conditions and provided that the marginal mean model specification is correct, T_S is chi-square distributed with r degrees of freedom under H_0.

Due to similar computational instability as for the Wald tests, Rotnizky and Jewell (1990) also propose the use of the 'working' score statistic T_S^*.

The 'Working' Score Test Statistic

The 'working' score test statistic T_S^* is defined by

$$T_S^* = \mathbf{U}^{(1)}\left(\boldsymbol{\beta}^0, \widehat{\boldsymbol{\beta}}^{(2)}\right)\left[\widehat{\text{Var}\left(\widehat{\boldsymbol{\beta}}^{(1)}\right)_M}\right]_{\boldsymbol{\beta}=\left(\boldsymbol{\beta}^0,\widehat{\boldsymbol{\beta}}^{(2)}\right)}\mathbf{U}^{(1)}\left(\boldsymbol{\beta}^0, \widehat{\boldsymbol{\beta}}^{(2)}\right), \tag{2.88}$$

where $\left[\widehat{\text{Var}\left(\widehat{\boldsymbol{\beta}}^{(1)}\right)_M}\right]_{\boldsymbol{\beta}=\left(\boldsymbol{\beta}^0,\widehat{\boldsymbol{\beta}}^{(2)}\right)}$ is the $(r \times r)$ submatrix of the model-based estimator for $\widehat{\boldsymbol{\beta}}$, evaluated under H_0.

The distribution of the statistic T_S^* is also presented in Rotnizky and Jewell (1990). Rotnizky and Jewell (1990, Theorem 2) show that

$$T_S^* = T_W^* + o_p(1). \tag{2.89}$$

2.9.2 Testing Quality of Fit using Generalized Wedderburn Estimating Equations

The generalized Wedderburn estimating equations proposed in this thesis, fall under a GEE framework. In a typical GEE analysis, as mentioned in the previous section, once the model parameters are estimated, use is made of measures such as the QIC proposed by Pan (2001a) or an adaptation of it, so as to choose the best working correlation structure amongst a number of candidates. The QIC may also be used to check which explanatory variables are to be retained in the model.

In Section 2.4.4, it has been shown that the estimates obtained using the generalized Wedderburn approach are invariant to the values of the dispersion and correlation parameters. Thus, the steps that are undertaken to test the quality of fit of a model obtained under the generalized Wedderburn approach, will not involve choosing the best working correlation structure. Focus is directed towards checking which model coefficients are significant and towards checking the resulting residuals. Details on the former are presented in the subsequent section. The Pearson residual has already been presented in Section 2.8.2. In Section 3.6, a distance measure which may be used to check the goodness of fit of a model fitted to compositional data and which is based on the Pearson residual is proposed. The reason for showing further material related to the Pearson residual in Section 3.6 is that Chapter 3 gives detail about the method developed by Aitchison (1982, 1986) to model compositional response data and provides some formal similarities in Aitchison's method and the generalized Wedderburn method. The residuals and distance measures proposed for use under the generalized Wedderburn approach will thus be presented in Section 3.6 in relation to the residuals and distance measure used under Aitchison's approach.

2.9.2.1 Testing Quality of Fit of Nested Models

Testing for goodness of fit of nested models under the generalized Wedderburn approach may be carried out using the 'working' Wald statistic T_W^* and the 'working' score statistic T_S^* developed by Rotnizky and Jewell (1990). As mentioned earlier, a score test is preferred over a Wald test due to the score test being invariant to nonlinear transformations of the model. Since identification of the model coefficients of the multivariate logit model relies on choosing a set of contrasts between the coefficients, the invariance to reparametrization of a score test is a very important and appealing property. Typically though, linear transformations of the model coefficients are required for testing for significance of model coefficients in the multivariate logit model, making both Wald and score tests suitable to test the quality of fit of the multivariate logit model. The statistics T_W^* and T_S^* are chosen over T_W and T_S due to their computational simplicity, as mentioned in Rotnizky and Jewell (1990), and also due to the availability of the model-based variance estimator (2.67) which has been developed to cater for the specific variability in compositional variables.

Under the generalized Wedderburn approach, the null hypothesis being tested is H_0 : $\boldsymbol{\gamma}^{(1)} = \boldsymbol{\gamma}^0$ where $\boldsymbol{\gamma}^{(1)}$ is a subset of the $(J-1)(p+1)$-vector of model coefficients $\boldsymbol{\gamma}$ and $\boldsymbol{\gamma}^0$ denotes the vector of values assigned to $\boldsymbol{\gamma}^{(1)}$ under H_0.

The 'Working' Wald Statistic T_W^* under the Generalized Wedderburn Approach

Under the assumption that the working correlation structure is the true correlation structure and under the generalized Wedderburn approach, the statistic T_W^* is given by

$$T_W^* = \left(\widehat{\boldsymbol{\gamma}}^{(1)} - \boldsymbol{\gamma}^0\right)' \left[\widehat{\text{Var}\left(\widehat{\boldsymbol{\gamma}}^{(1)}\right)}_M\right]^{-1} \left(\widehat{\boldsymbol{\gamma}}^{(1)} - \boldsymbol{\gamma}^0\right) \tag{2.90}$$

where $\widehat{\text{Var}\left(\widehat{\boldsymbol{\gamma}}^{(1)}\right)}_M$ is the $(r \times r)$ submatrix of the model-based estimator (2.67) for $\widehat{\boldsymbol{\gamma}}$.
As per Rotnizky and Jewell (1990), if the chosen working correlation structure is the true correlation structure, T_W^* is asymptotically chi-square distributed with r degrees of freedom.

The 'Working' Score Statistic T_S^* under the Generalized Wedderburn Approach

Using (2.54), the estimating functions for $\boldsymbol{\gamma}$, in conjunction with the working variance-

covariance matrix $\phi\mathbb{V}_{\mathbf{p}_i,\boldsymbol{\Omega},\mathbb{W}} = \left(\mathbb{P}_i - \mathbf{p}_i\mathbf{p}_i{}'\right)\boldsymbol{\Sigma}\left(\mathbb{P}_i - \mathbf{p}_i\mathbf{p}_i{}'\right)$, T_S^* becomes

$$T_S^* = \left[\sum_{i=1}^{n}\mathbb{X}_i'\left(\widehat{\mathbb{P}}_i - \widehat{\mathbf{p}}_i\widehat{\mathbf{p}}_i'\right)\left[\hat{\phi}\widehat{\mathbb{V}_{\mathbf{p}_i,\boldsymbol{\Omega},\mathbb{W}}}\right]^{-}(\mathbf{Y}_i - \widehat{\mathbf{p}}_i)\right]' \left[\widehat{\mathrm{Var}\left(\widehat{\boldsymbol{\gamma}}^{(1)}\right)}_M\right]_{\boldsymbol{\gamma}=\left(\boldsymbol{\gamma}^0,\widehat{\boldsymbol{\gamma}}^{(2)}\right)} \times \left[\sum_{i=1}^{n}\mathbb{X}_i'\left(\widehat{\mathbb{P}}_i - \widehat{\mathbf{p}}_i\widehat{\mathbf{p}}_i'\right)\left[\hat{\phi}\widehat{\mathbb{V}_{\mathbf{p}_i,\boldsymbol{\Omega},\mathbb{W}}}\right]^{-}(\mathbf{Y}_i - \widehat{\mathbf{p}}_i)\right] \tag{2.91}$$

where $\left[\widehat{\mathrm{Var}\left(\widehat{\boldsymbol{\gamma}}^{(1)}\right)}_M\right]_{\boldsymbol{\gamma}=\left(\boldsymbol{\gamma}^0,\widehat{\boldsymbol{\gamma}}^{(2)}\right)}$ is the $(r \times r)$ submatrix of the model-based estimator (2.67) evaluated under H_0 and $\phi\mathbb{V}_{\mathbf{p}_i,\boldsymbol{\Omega},\mathbb{W}}$ is estimated as described in Section 2.8.2.

Analogous to the 'working' Wald statistic, as per Rotnizky and Jewell (1990), if the chosen working correlation structure is the true correlation structure, T_S^* is asymptotically chi-square distributed with r degrees of freedom.

2.9.2.2 The Non-Suitability of Pan's QIC Criterion

Despite its popularity and ease of implementation, Pan's QIC may not be used under the generalized Wedderburn approach. As seen in equation (2.83), Pan's QIC relies on the specification of a log quasi-likelihood function. In this thesis, an attempt at obtaining a log quasi-likelihood function suitable for use with the generalized Wedderburn estimating equations is made. Details of why it is impossible to obtain such a function, are presented shortly.

Details on the Non-Existence of a Log Quasi-Likelihood Function

Let us first consider the case $J = 2$. We will show that this case is special, in that there *does* exist a log quasi-likelihood function when $J = 2$. Unfortunately though, as will be shown, this existence property does not extend to values of J greater than 2. In the special case $J = 2$, we will show here that the Wedderburn estimating equations minimize a deviance function that can be written intriguingly as a linear combination of the familiar gamma and Poisson deviance functions. This turns out, though, to be more of an interesting curiosity than a route to a more general result.

Consider the deviance functions $d_G(y_{ij}, \hat{p}_{ij})$ and $d_P(y_{ij}, \hat{p}_{ij})$, corresponding to the gamma and Poisson distribution respectively, with $\hat{p}_{ij}$ being the estimator of p_{ij}. So

$$d_G(y_{ij}, \hat{p}_{ij}) = 2\left[-\log\left(\frac{y_{ij}}{\hat{p}_{ij}}\right) + \frac{(y_{ij} - \hat{p}_{ij})}{\hat{p}_{ij}}\right] \tag{2.92}$$

and

$$d_P(y_{ij}, \hat{p}_{ij}) = 2\left[y_{ij}\log\left(\frac{y_{ij}}{\hat{p}_{ij}}\right) - (y_{ij} - \hat{p}_{ij})\right]. \tag{2.93}$$

Also consider the function $d\left(y_{ij}, p_{ij}\right)$ defined by

$$d\left(y_{ij}, p_{ij}\right) = d_G\left(y_{ij}, p_{ij}\right) + 2d_P\left(y_{ij}, p_{ij}\right).$$

For case i, let

$$d^+\left(\mathbf{y}_i, \mathbf{p}_i\right) = \sum_{j=1}^{2} d\left(y_{ij}, p_{ij}\right),$$

so that

$$d^+\left(\mathbf{y}_i, \mathbf{p}_i\right) = d\left(y_{i1}, p_{i1}\right) + d\left(1 - y_{i1}, 1 - p_{i1}\right).$$

Differentiating $d^+\left(\mathbf{y}_i, \mathbf{p}_i\right)$ with respect to p_{i1} leads to

$$\frac{\partial d^+}{\partial p_{i1}} = -2\left(\frac{y_{i1} - p_{i1}}{p_{i1}^2\left(1 - p_{i1}\right)^2}\right), \tag{2.94}$$

which is $-2\times$ the quasi-score function corresponding to a response variable Y_{i1} with mean p_{i1} and a mean-variance relationship defined by $V(p_{i1}) = p_{i1}^2\left(1 - p_{i1}\right)^2$. So the function $\sum_{i=1}^{n} d^+\left(\mathbf{y}_i, \mathbf{p}_i\right)$ may actually serve as a log quasi-likelihood function for Wedderburn's (1974) estimating equations.

For $J = 2$, it may also be shown that the Pearson chi-square statistic for the gamma distribution added to twice the Pearson chi-square statistic for the Poisson distribution is the same as to the Pearson chi-square statistic achieved under the mean-variance specification used in Wedderburn (1974). Details follow:

$$\begin{aligned}\sum_{i=1}^{n}\sum_{j=1}^{2}\left[\frac{\left(y_{ij} - p_{ij}\right)^2}{p_{ij}^2} + 2\frac{\left(y_{ij} - p_{ij}\right)^2}{p_{ij}}\right] &= \sum_{i=1}^{n}\left[\frac{\left(y_{i1} - p_{i1}\right)^2}{p_{i1}^2} + \frac{\left(1 - y_{i1} - \left(1 - p_{i1}\right)\right)^2}{\left(1 - p_{i1}\right)^2}\right. \\ &\quad \left. +2\left(\frac{\left(y_{i1} - p_{i1}\right)^2}{p_{i1}} + \frac{\left(1 - y_{i1} - \left(1 - p_{i1}\right)\right)^2}{1 - p_{i1}}\right)\right] \\ &= \sum_{i=1}^{n}\frac{\left(y_{i1} - p_{i1}\right)^2}{p_{i1}^2\left(1 - p_{i1}\right)^2}.\end{aligned} \tag{2.95}$$

The relationships relating the deviance functions and Pearson chi-square statistics for the gamma and Poisson distribution to Wedderburn's (1974) mean-variance specification, do not however generalize to cater for $J > 2$. On using the estimating function U_{js}, defined in (2.55), for $j \neq j^{'}$ and $j, j^{'} = 1, \ldots, J-1$, it may be shown that

$$\frac{\partial U_{js}}{\partial \gamma_{j^{'}k}} \neq \frac{\partial U_{j^{'}k}}{\partial \gamma_{js}}, \tag{2.96}$$

showing that the matrix of derivatives is not symmetric. For the generalized Wedderburn estimating equations, the quasi log-likelihood function is thus not uniquely defined. The proof showing that the matrix of derivatives is not symmetric is presented in Appendix B.

Chapter 3

The Relationship between the Generalized Wedderburn Model and Aitchison (1982, 1986) Regression Model

3.1 Introduction

As mentioned in Section 1.2.2, the standard methodology used to model compositional response variables is that devised by Aitchison (1982, 1986), which involves taking one compositional variable as reference component, taking the logratios of the response variables with respect to the chosen reference component and then proceeding to model the influence of the explanatory variables on the transformed response variables using linear regression analysis. If no zeros are present in the compositional variables and with the assumption that the logratios are multivariate normally distributed, the model coefficients are estimated using maximum likelihood estimation. Maximum likelihood estimators are well understood and have desirable properties under the model.

Whilst the generalized Wedderburn approach models compositional data on the original scale through the multivariate logit model, in Section 3.3 it will be shown how the approach developed by Aitchison (1982, 1986) may be viewed as an additive model which has been obtained as a result of taking the logarithm on both sides of the multiplicative model (2.1). Although the two approaches are both related to the multiplicative model (2.1), the two approaches estimate different mean models. Detail of why the two approaches estimate different mean models is presented in Section 3.4. The two approaches do however have some striking similarities of form. The formal similarities of the two approaches are presented in Section 3.5. Section 3.6 presents residuals and a distance measure (Aitchison, 1992) that may be used to check the goodness of fit of a model fitted to compositional

data. A distance measure based on the Pearson residuals achieved under the generalized Wedderburn approach is also developed. An in-depth study of the properties of estimators obtained under the two methods is then made in Section 3.7.

Due to the fact that the generalized Wedderburn method and Aitchison's method estimate different mean models, efficiency comparisons between the estimators used to estimate the model coefficients do not in general make sense. A comparison of efficiency between the two estimators may however be undertaken if the true values of the model coefficients corresponding to the explanatory variables are set equal to zero in both models simultaneously. Such a null-effects situation will be considered in Section 3.8 where a small simulation study is carried out to compare the efficiency of the GEE estimator to that of the MLE under various sample sizes, coefficients of variation and correlation coefficients, with compositional data being generated through multivariate lognormally distributed $\dot{\mathbf{Y}}_i$. Under the generalized Wedderburn approach, two different estimates of the variance-covariance matrix $\text{Var}\left(\widehat{\boldsymbol{\gamma}}\right)$ are also obtained for each sample; the model-based estimator (2.67) with $\hat{\phi}\widehat{\mathbb{V}_{\mathbf{p}_i,\boldsymbol{\Omega},\mathbb{W}}}$ computed using (2.74) and the robust estimator of Liang and Zeger (1986) described in Section 2.8.3. This is done in order to be able to compare the performance of two estimators under various sample sizes, coefficients of variation and correlation coefficients.

The generalized Wedderburn method and Aitchison's method are then compared on two widely used datasets from the compositional data literature, the Arctic Lake dataset (e.g. Aitchison, 1986; Tsagris et al., 2011; Maier, 2014) and the Foraminiferal dataset (e.g. Aitchison, 1986; Palarea-Albaladejo et al., 2007; Scealy and Welsh, 2011; Tsagris, 2015), in Sections 3.9 and 3.10 respectively.

3.2 The Multiplicative Regression Model (MRM)

Recall that for a sample of size n and a set of predictors $X_1, \ldots, X_p$, the random variables $\dot{Y}_{ij}$, $i = 1, \ldots, n$, $j = 1, \ldots, J$ are modeled multiplicatively as

$$\dot{Y}_{ij} = m_i\left(\dot{\theta}_i, \boldsymbol{\beta}_j\right) E_{ij}, \tag{3.1}$$

where the function m_i for the i^{th} case is defined as

$$m_i\left(\theta, \boldsymbol{\beta}\right) = \exp\left(\theta + \mathbf{x}_i^{'}\boldsymbol{\beta}\right), \tag{3.2}$$

the error vectors $\mathbf{E}_i = \left(E_{i1}, \ldots, E_{iJ}\right)^{'}$ are assumed to be independent of one another, $E\left(\mathbf{E}_i\right) = \mathbf{1}$ and $\text{Var}\left(\mathbf{E}_i\right) = \dot{\boldsymbol{\Sigma}} = \phi\boldsymbol{\Omega}^{\frac{1}{2}}\mathbb{W}\boldsymbol{\Omega}^{\frac{1}{2}}$ with ϕ being a common dispersion parameter, $\boldsymbol{\Omega} = \text{diag}\left(\omega_1, \ldots, \omega_J\right)$ with $\omega_1, \ldots, \omega_J$ being relative dispersion parameters, and $\mathbb{W}$ is an unknown correlation matrix.

If $\alpha_{jj'}$ denotes the typical element in the $J \times J$ correlation matrix $\mathbb{W}$, it follows that

$$\begin{aligned} E\left(\dot{Y}_{ij}\right) &= m_i\left(\dot{\theta}_i, \boldsymbol{\beta}_j\right), \\ \text{Var}\left(\dot{Y}_{ij}\right) &= \phi\omega_j\left[m_i\left(\dot{\theta}_i, \boldsymbol{\beta}_j\right)\right]^2, \\ \text{Cov}\left(\dot{Y}_{ij}, \dot{Y}_{ij'}\right) &= \phi\sqrt{\omega_j}\sqrt{\omega_{j'}} m_i\left(\dot{\theta}_i, \boldsymbol{\beta}_j\right) m_i\left(\dot{\theta}_i, \boldsymbol{\beta}_{j'}\right)\alpha_{jj'}. \end{aligned} \tag{3.3}$$

The multiplicative model (3.1) motivates two distinct ways of estimating the model parameters. One approach models the response variables on the original scale through the multiplicative model whilst the other approach makes use of the additive model that is obtained as a result of taking the logarithm on both sides of the multiplicative model.

For compositional response variables, in the previous chapter we have seen how the first approach leads to the generalized Wedderburn method where $E(Y_{ij})$ is modeled using a multivariate logit model. The second approach leads to Aitchison's regression analysis of logratios, modeling $E(\log(Y_{ij}/Y_{iJ}))$, if component J is taken as the reference component. A short introduction on Aitchison's logratio method has been given in Section 1.2.2. More details on Aitchison's method and the way this approach relates to the multiplicative regression model (3.1) and the generalized Wedderburn method, are given in Sections 3.3, 3.4 and 3.5.

3.3 Aitchison (1982, 1986) Regression Model

Consider the following definition which gives the details upon which the Aitchison (1982, 1986) modeling strategy for compositional data is based.

Definition 3.3.1. *(Aitchison, 1986) Let $\mathbf{Y}_i$ be a J-part compositional vector for case i and suppose that $\mathbf{W}_i = \left(\log\left(\frac{Y_{i1}}{Y_{iJ}}\right), \ldots, \log\left(\frac{Y_{i,J-1}}{Y_{iJ}}\right)\right)' = \mathbb{F}\log(\mathbf{Y}_i)$ follows a multivariate normal distribution with mean vector $\mathbb{F}\boldsymbol{\zeta}_i$ and variance covariance matrix $\mathbb{F}\boldsymbol{\Psi}\mathbb{F}'$, where $\mathbb{F}$ is the $(J-1)\times J$ matrix $[\mathbb{I}_{J-1}, -\mathbf{1}_{J-1}]$, $\mathbb{I}_{J-1}$ being a $(J-1)\times(J-1)$ identity matrix and $\mathbf{1}_{J-1}$ being a $(J-1)$-vector of ones. Then*

1. *the J-vector of latent variables $\dot{\mathbf{Y}}_i$, in the positive space $\mathbb{R}_+^J$, follows a multivariate lognormal distribution with parameters $\boldsymbol{\zeta}_i$ and $\boldsymbol{\Psi}$,*
2. *the J-part composition $\mathbf{Y}_i$ follows an additive logistic normal distribution with parameters $\mathbb{F}\boldsymbol{\zeta}_i$ and $\mathbb{F}\boldsymbol{\Psi}\mathbb{F}'$.*

If as per Definition 3.3.1, $\dot{\mathbf{Y}}_i$ follows a multivariate lognormal distribution with parameters $\boldsymbol{\zeta}_i$ and $\boldsymbol{\Psi}$, then, $\log\left(\dot{\mathbf{Y}}_i\right)$ follows a multivariate normal distribution with mean vector $\boldsymbol{\zeta}_i$ and variance covariance matrix $\boldsymbol{\Psi}$. Letting the diagonal elements of $\boldsymbol{\Psi}$ be denoted by ψ_j^2

and its off-diagonal elements be denoted by $\psi_{jj'}$, $\left(j, j^{'} = 1, \ldots, J\right)$, $j \neq j^{'}$, the variance-covariance matrix $\mathbb{F}\boldsymbol{\Psi}\mathbb{F}^{'}$ is given by

$$\mathbb{F}\boldsymbol{\Psi}\mathbb{F}^{'} = \begin{pmatrix} \psi_1^2 - 2\psi_{1J} + \psi_J^2 & \ldots & \psi_{1,J-1} - \psi_{1,J} - \psi_{J-1,J} + \psi_J^2 \\ & \ddots & \vdots \\ & & \psi_{J-1}^2 - 2\psi_{J-1,J} + \psi_J^2 \end{pmatrix} \tag{3.4}$$

where $\mathbb{F} = \begin{pmatrix} 1 & 0 & \ldots & 0 & -1 \\ \vdots & \ddots & & \vdots & \\ 0 & \ldots & 0 & 1 & -1 \end{pmatrix}$.

Additionally, if $\boldsymbol{\zeta}_i = (\zeta_{i1}, \ldots, \zeta_{iJ})^{'}$, then

$$\begin{aligned} E\left(\dot{Y}_{ij}\right) &= \exp\left(\zeta_{ij} + \frac{1}{2}\psi_j^2\right), \\ \text{Var}\left(\dot{Y}_{ij}\right) &= \left(\exp\left(\psi_j^2\right) - 1\right)\exp\left(2\zeta_{ij} + \psi_j^2\right), \end{aligned} \tag{3.5}$$

and through the use of the moment generating function of the multivariate normal distribution,

$$\text{Cov}\left(\dot{Y}_{ij}, \dot{Y}_{ij'}\right) = m_i\left(\dot{\theta}_i, \boldsymbol{\beta}_j\right) m_i\left(\dot{\theta}_i, \boldsymbol{\beta}_{j'}\right)\left(\exp\left(\psi_{jj'}\right) - 1\right). \tag{3.6}$$

Then, on relating equations (3.5) and (3.6) with (3.3), the latter obtained under the latent multiplicative regression model, we have that

$$\zeta_{ij} = \log\left(m_i\left(\dot{\theta}_i, \boldsymbol{\beta}_j\right)\right) - \frac{1}{2}\psi_j^2, \tag{3.7}$$

$$\phi\omega_j = \exp(\psi_j^2) - 1, \tag{3.8}$$

$$\psi_j^2 = \log(\phi\omega_j + 1) \tag{3.9}$$

and

$$\psi_{jj'} = \log\left(\phi_j\sqrt{\omega_j}\sqrt{\omega_{j'}}\alpha_{jj'} + 1\right). \tag{3.10}$$

Also, as per Definition 3.3.1, the vector of logratios $\mathbf{W}_i = \left(\log\left(\frac{Y_{i1}}{Y_{iJ}}\right), \ldots, \log\left(\frac{Y_{i,J-1}}{Y_{iJ}}\right)\right)^{'}$ is assumed to follow a multivariate normal distribution with mean vector $\mathbb{F}\boldsymbol{\zeta}_i$ and variance-covariance matrix $\mathbb{F}\boldsymbol{\Psi}\mathbb{F}^{'}$. Using (3.7) and (3.2), the components making up $\mathbb{F}\boldsymbol{\zeta}_i$ are given

by

$$
\begin{aligned}
E\left[\log\left(\frac{Y_{ij}}{Y_{iJ}}\right)\right] &= \zeta_{ij} - \zeta_{iJ} \\
&= \left(\log\left(m_i\left(\dot{\theta}_i, \boldsymbol{\beta}_j\right)\right) - \frac{1}{2}\psi_j^2\right) - \left(\log\left(m_i\left(\dot{\theta}_i, \boldsymbol{\beta}_J\right)\right) - \frac{1}{2}\psi_J^2\right) \\
&= \left(\dot{\theta}_i + \mathbf{x}_i^{'}\boldsymbol{\beta}_j - \frac{1}{2}\log(\phi\omega_j + 1)\right) - \left(\dot{\theta}_i + \mathbf{x}_i^{'}\boldsymbol{\beta}_J - \frac{1}{2}\log(\phi\omega_J + 1)\right) \\
&= \left(\dot{\theta}_i + \mathbf{x}_i^{'}\boldsymbol{\beta}_j^*\right) - \left(\dot{\theta}_i + \mathbf{x}_i^{'}\boldsymbol{\beta}_J^*\right)
\end{aligned}
$$

where, for $j = 1, \ldots, J$, $\boldsymbol{\beta}_j^* = \left(\beta_{j0}^*, \beta_{j1}, \ldots, \beta_{jp}\right)^{'}$ and $\beta_{j0}^* = \beta_{j0} - \frac{1}{2}\log(\phi\omega_j + 1)$.

Consequently

$$
\begin{aligned}
E\left(\log\left(\dot{Y}_{ij}\right)\right) - E\left(\log\left(\dot{Y}_{iJ}\right)\right) &= E\left(\log\left(Y_{ij}\right)\right) - E\left(\log\left(Y_{iJ}\right)\right) \\
&= \left(\dot{\theta}_i + \mathbf{x}_i^{'}\boldsymbol{\beta}_j^*\right) - \left(\dot{\theta}_i + \mathbf{x}_i^{'}\boldsymbol{\beta}_J^*\right) \\
&= \mathbf{x}_i^{'}\left(\boldsymbol{\beta}_j^* - \boldsymbol{\beta}_J^*\right) \\
&= \mathbf{x}_i^{'}\boldsymbol{\gamma}_j^*,
\end{aligned} \tag{3.11}
$$

where $\boldsymbol{\gamma}_j^* = \left(\gamma_{j0}^*, \gamma_{j1}, \ldots, \gamma_{jp}\right)^{'}$ with $\gamma_{j0}^* = \gamma_{j0} - \frac{1}{2}\log\left(\frac{\phi\omega_j+1}{\phi\omega_J+1}\right)$.

The steps in (3.11) explain the relationship that is modeled when using Aitchison's logratio method. Except for a change in notation for the model coefficients, the equation which results from (3.11) is exactly the same as equation (1.8). In Chapter 1, Aitchison's method has been presented in terms of $\boldsymbol{\beta}$ as unless otherwise stated, throughout the thesis, $\boldsymbol{\beta}$ is the notation that is used to represent a general vector of model coefficients.

The steps in (3.11) also lead to

$$
E\left(\log\left(\dot{Y}_{ij}\right)\right) = \dot{\theta}_i + \mathbf{x}_i^{'}\boldsymbol{\beta}_j^*. \tag{3.12}
$$

Based on (3.11) and (3.12) it may thus be noticed that, given some sample data, Aitchison's method may be viewed as being the result of fitting a standard linear model to each log-transformed compositional variable (or equivalently to each log-transformed latent variable $\dot{Y}_{ij}$), choosing a reference component and thus proceeding to take the differences in the resulting estimates with respect to the chosen reference component.

More specifically, Aitchison's method is based on taking logs on both sides of equation

(3.1). This leads to

$$\begin{aligned}\log\left(\dot{Y}_{ij}\right) &= \log\left(m_i\left(\dot{\theta}_i, \boldsymbol{\beta}_j\right)\right) + \log\left(E_{ij}\right) \\ &= \log\left(m_i\left(\dot{\theta}_i, \boldsymbol{\beta}_j\right)\right) + \log\left(E_{ij}\right) - E\left(\log\left(E_{ij}\right)\right) + E\left(\log\left(E_{ij}\right)\right) \\ &= \log\left(m_i\left(\dot{\theta}_i, \boldsymbol{\beta}_j\right)\right) + E\left(\log\left(E_{ij}\right)\right) + E_{ij}^{*}\end{aligned} \tag{3.13}$$

where $E_{ij}^{*} = \log\left(E_{ij}\right) - E\left(\log\left(E_{ij}\right)\right)$. Clearly $E\left(E_{ij}^{*}\right) = 0$. Assume that E_{ij}^{*} are independent and identically distributed amongst cases i for every component j. Then

$$\begin{aligned}E\left(\log\left(\dot{Y}_{ij}\right)\right) &= \log\left(m_i\left(\dot{\theta}_i, \boldsymbol{\beta}_j\right)\right) + E\left(\log\left(E_{ij}\right)\right) \\ &= \dot{\theta}_i + \left[\beta_{j0} + E\left(\log\left(E_{ij}\right)\right)\right] x_{i0} + \beta_{j1} x_{i1} + \cdots + \beta_{jp} x_{ip} \\ &= \dot{\theta}_i + \mathbf{x}_i^{'} \boldsymbol{\beta}_j^{*}\end{aligned} \tag{3.14}$$

which is the same as equation (3.12).

So whilst the generalized Wedderburn method models compositional data on the original scale through the multivariate logit model, the method developed by Aitchison (1982, 1986) may be viewed as an additive model which has been obtained as a result of taking the logarithm of the multiplicative model. One point worth mentioning at this stage is that despite the correspondence of the two approaches with the multiplicative regression model, the two approaches are based on different mean models which lead to different parameter estimates. Also, although Aitchison (1982, 1986) associates the modeling of the logratios with the multivariate normal distribution, Aitchison's approach may actually be used on any set of logratios, irrespective of their distribution. This is because, on using Aitchison's approach, estimation of the model parameters is carried out by means of the method of ordinary least squares. So, unless otherwise specified, the discussion about Aitchison's method which follows, will be more general than that presented by Aitchison (1982, 1986). More detail is provided in the following sections.

3.4 The Differences between the Generalized Wedderburn Model and Aitchison (1982, 1986) Model

In general, estimation of the parameters of a multiplicative model is carried out by either considering the model for the response variable on the original scale or by otherwise considering the model based on a log-transformed response variable. Except for the intercept, the two methods estimate the same set of parameters (e.g. Firth, 1988). Despite being related to the same multiplicative model, the generalized Wedderburn method and Aitchison's method are however two different methods which estimate two different sets of model parameters.

More specifically, in Section 2.7 it was shown that under the generalized Wedderburn approach,

$$E\left(Y_{ij}\right) = E\left(\frac{\dot{Y}_{ij}}{\dot{Y}_{i1} + \cdots + \dot{Y}_{iJ}}\right) \approx p_{ij} = \frac{E\left(\dot{Y}_{ij}\right)}{E\left(\dot{Y}_{i1}\right) + \cdots + E\left(\dot{Y}_{iJ}\right)}, \tag{3.15}$$

and in the generalized Wedderburn estimating equations (2.55), the relationship between $E\left(Y_{ij}\right)$ and p_{ij} is taken to be exact rather than approximate as in (3.15).

Under the generalized Wedderburn approach

$$E\left(Y_{ij}\right) = p_{ij} = \frac{m_i\left(\dot{\theta}_i, \boldsymbol{\beta}_j\right)}{\sum_{j'=1}^{J} m_i\left(\dot{\theta}_i, \boldsymbol{\beta}_{j'}\right)} = \frac{\exp\left(\mathbf{x}_i^{'}\boldsymbol{\beta}_j\right)}{\sum_{j'=1}^{J} \exp\left(\mathbf{x}_i^{'}\boldsymbol{\beta}_{j'}\right)}, \tag{3.16}$$

which, under reparametrization, leads to

$$E\left(Y_{ij}\right) = p_{ij} = \frac{\exp\left(\mathbf{x}_i^{'}\boldsymbol{\gamma}_j\right)}{\sum_{j'=1}^{J} \exp\left(\mathbf{x}_i^{'}\boldsymbol{\gamma}_{j'}\right)}. \tag{3.17}$$

The γ estimates obtained through (3.17) are thus different from the estimates obtained for the corresponding parameters in Aitchison's regression model. In view of the Taylor Series approximation, the differences between the two sets of parameters (except for the intercept) only diminish as the dispersion ϕ goes to zero.

3.5 The Formal Similarities between the Generalized Wedderburn Model and Aitchison (1982, 1986) Model

Despite being two different methods which estimate two different mean models, Aitchison's approach and the generalized Wedderburn approach share a number of similarities of form.

3.5.1 The Generalized Wedderburn Model

The generalized Wedderburn approach has been developed by specifying the mean-variance relationship of latent variable $\dot{Y}_{ij}$ to be as for a family of gamma distributions with constant coefficient of variation together with a log link function to model $\log\left(E\left(\dot{Y}_{ij}\right)\right) = \theta_i + \mathbf{x}_i^{'}\boldsymbol{\beta}_j$ directly. In Section 2.4.2, it has also been shown that the estimating equations for $\boldsymbol{\beta}_1, \ldots, \boldsymbol{\beta}_J$ are given by

$$\sum_{i=1}^{n}\left(\frac{Y_{ij}}{m_i\left(\theta_i,\boldsymbol{\beta}_j\right)}-1\right)x_{i,s-1}=0 \qquad \text{for } j=1,\ldots,J, \quad s=1,\ldots,p+1. \tag{3.18}$$

Note that the same γ estimates will be achieved if either estimating equations (3.18) or estimating equations (2.55) are used under the generalized Wedderburn approach. If estimating equations (3.18) are used, the γ estimates are achieved through an iterative process which involves the estimation of both $\boldsymbol{\theta}$ and $\boldsymbol{\beta}$, until convergence is achieved. If estimating equations (2.55) are used, an iterative process still has to be used but no θ parameters need to be estimated. Focus here is directed towards estimating equations (3.18) due to the similarities that will be drawn between this set of estimating equations and the estimating equations that are used in Aitchison's method.

On using estimating equation (3.18), if $\widehat{\boldsymbol{\beta}}$ is the estimator of $\boldsymbol{\beta}=\left(\boldsymbol{\beta}_1',\ldots,\boldsymbol{\beta}_J'\right)'$, in Section 2.4.4 we have seen that $\widehat{\boldsymbol{\beta}}$ is the ordinary least squares estimator

$$\widehat{\boldsymbol{\beta}}=\left[\mathbb{I}_J\otimes\left(\mathbb{X}'\mathbb{X}\right)^{-1}\mathbb{X}'\right]\mathbf{Z}^*, \tag{3.19}$$

where $\mathbb{I}_J$ is a $J\times J$ identity matrix, $\mathbb{X}$ is the design matrix and $\mathbf{Z}^*=\left(\mathbf{Z}_{(1)}^{*\prime},\ldots,\mathbf{Z}_{(J)}^{*\prime}\right)'$ is the vector of working variates. For $i=1,\ldots,n, j=1,\ldots,J$, $\mathbf{Z}_{(j)}^*=\left(Z_{1j}^*,\ldots,Z_{nj}^*\right)'$ where

$$Z_{ij}^*=\log\left(m_i\left(\widehat{\theta}_i,\widehat{\boldsymbol{\beta}}_j\right)\right)+\left(\frac{Y_{ij}}{m_i\left(\widehat{\theta}_i,\widehat{\boldsymbol{\beta}}_j\right)}-1\right). \tag{3.20}$$

In the equation just provided, $m_i\left(\widehat{\theta}_i,\widehat{\boldsymbol{\beta}}_j\right)$ denotes the fitted value of $m_i\left(\theta_i,\boldsymbol{\beta}_j\right)$. It should be noted that at convergence $m_i\left(\widehat{\theta}_i,\widehat{\boldsymbol{\beta}}_j\right)=\hat{p}_{ij}$ where $\hat{p}_{ij}$ is the estimator of p_{ij} (see equation (3.16)).

3.5.2 Aitchison's Model

The estimating equations which correspond to (3.14) are given by

$$\sum_{i=1}^{n}\left[\log\left(\dot{Y}_{ij}\right)-\log\left(m_i\left(\dot{\theta}_i,\boldsymbol{\beta}_j^*\right)\right)\right]x_{i,s-1}=0 \qquad \text{for } j=1,\ldots,J, \quad s=1,\ldots,p+1 \tag{3.21}$$

where $\log\left(m_i\left(\dot{\theta}_i,\boldsymbol{\beta}_j^*\right)\right)=\log\left(m_i\left(\dot{\theta}_i,\boldsymbol{\beta}_j\right)\right)+E\left(\log\left(E_{ij}\right)\right)$.

Since $\log\left(\dot{Y}_{ij}\right)-\log\left(m_i\left(\dot{\theta}_i,\boldsymbol{\beta}_j^*\right)\right)=\log\left(Y_{ij}\right)-\log\left(m_i\left(\theta_i,\boldsymbol{\beta}_j^*\right)\right)$, estimating equations (3.21) may be rewritten as

$$\sum_{i=1}^{n}\left[\log\left(Y_{ij}\right)-\log\left(m_i\left(\theta_i,\boldsymbol{\beta}_j^*\right)\right)\right]x_{i,s-1}=0 \qquad \text{for } j=1,\ldots,J, \quad s=1,\ldots,p+1, \tag{3.22}$$

where (3.22) are ordinary least squares equations for compositional variables Y_{ij}. Under the assumption that $\log(Y_{ij})$ is normally distributed with mean $\log\left(m_i\left(\theta_i, \boldsymbol{\beta}_j^*\right)\right)$ and with constant variance across all cases i for each component j, the estimating equations (3.22) may also be considered to be maximum likelihood equations.

In contrast with the generalized Wedderburn approach, Aitchison's approach delivers the estimates of the β parameters in one evaluation step. If $\boldsymbol{\beta}^* = \left(\boldsymbol{\beta}_1^{*'}, \ldots, \boldsymbol{\beta}_J^{*'}\right)'$, its estimator $\widehat{\boldsymbol{\beta}^*}$, achieved under Aitchison's approach, is however also an ordinary least squares estimator. The estimator $\widehat{\boldsymbol{\beta}^*}$ is given by

$$\widehat{\boldsymbol{\beta}^*} = \left[\mathbb{I}_J \otimes \left(\mathbb{X}'\mathbb{X}\right)^{-1}\mathbb{X}'\right]\mathbf{Z}^*, \tag{3.23}$$

where $\mathbb{I}_J$ is a $J \times J$ identity matrix, $\mathbb{X}$ is the design matrix and $\mathbf{Z}^* = \left(\mathbf{Z}_{(1)}^{*'}, \ldots, \mathbf{Z}_{(J)}^{*'}\right)'$ is the vector of working variates. For $i = 1, \ldots, n, j = 1, \ldots, J$, $\mathbf{Z}_{(j)}^* = \left(Z_{1j}^*, \ldots, Z_{nj}^*\right)'$ where in this case,

$$Z_{ij}^* = \log(Y_{ij}). \tag{3.24}$$

A further similarity in the two methods will be noticed once the variance-covariance matrix of the γ parameters is derived, under the two approaches, in Section 3.7.

From the working variates (3.20) and (3.24), used under the two different methods, it may however be appreciated why, of these two approaches, only the generalized Wedderburn approach can still be used if any zeros are present in the compositional response variables. The generalized Wedderburn approach is also preferable for interpretation purposes, as this approach models $E(Y_{ij})$ directly.

3.6 Residuals and Distance Measures based on the Two Models

3.6.1 Residuals and Distance Measure based on Aitchison's Model

As seen in the previous section, estimates under Aitchison's approach are obtained using ordinary least squares. Thus, for $i = 1, \ldots, n, j = 1 \ldots, J$, the residuals are given by

$$\log(Y_{ij}) - \log\left(\hat{p}_{ij}^*\right), \tag{3.25}$$

where $\hat{p}_{ij}^*$ is the fitted value of Y_{ij}. The fitted values $\hat{p}_{ij}^*$ are obtained on exponentiating and rescaling the fitted values $m_i\left(\widehat{\theta}_i, \widehat{\boldsymbol{\beta}}_j^*\right)$ obtained from fitting a standard linear regression model to the components of $\log(\mathbf{Y})$.

Alternatively, a centred version of the residuals (3.25) may be used, giving

$$R_{ij}^{*} = \log\left(\frac{Y_{ij}}{\hat{p}_{ij}^{*}}\right) - \frac{1}{J}\sum_{j'=1}^{J}\log\left(\frac{Y_{ij'}}{\hat{p}_{ij'}^{*}}\right). \tag{3.26}$$

In analyzing the Arctic Lake dataset and the Foraminiferal dataset in Sections 3.9 and 3.10 respectively, use is made of the centred residual R_{ij}^{*}. Preference is given to the centred residuals as, for all j and s, these residuals satisfy

$$\sum_{i=1}^{n} R_{ij}^{*} x_{is} = 0, \tag{3.27}$$

and so they are orthogonal to the columns of the design matrix $\mathbb{X}$. Furthermore, these centred residuals will be shown to be directly related to the distance measure developed by Aitchison (1992). This distance measure may be used to check the goodness of fit of a model fitted to compositional data. Moreover, recall from Pg 50 that the Pearson residual used under the generalized Wedderburn approach also involves a centering operation and a distance measure based on Pearson residual will also be developed in Section 3.6.2.

As per Aitchison (1992), the distance between two compositions $\mathbf{Y}$ and $\mathbf{Y}^{*}$, $\Delta\left(\mathbf{Y}, \mathbf{Y}^{*}\right)$, should be measured by considering the difference in the logratios of the two compositions as follows:

$$\Delta\left(\mathbf{Y}, \mathbf{Y}^{*}\right) = \left[\sum_{j''=1}^{J}\underbrace{\sum_{j=1}^{J}}_{j<j''}\left[\log\left(\frac{Y_j}{Y_{j''}}\right) - \log\left(\frac{Y_j^{*}}{Y_{j''}^{*}}\right)\right]^2\right]^{\frac{1}{2}}. \tag{3.28}$$

The distance measure $\Delta\left(\mathbf{Y}, \mathbf{Y}^{*}\right)$ adheres to a number of criteria, which are line with the requirements of the geometrical structure of the simplex S^{J-1}. The following list is an adaptation of the criteria mentioned in Aitchison (1992, p. 374). The distance measure $\Delta\left(\mathbf{Y}, \mathbf{Y}^{*}\right)$

- is positive if compositions $\mathbf{Y}$ and $\mathbf{Y}^{*}$ are not equivalent
- is equal to zero if $\mathbf{Y} = a\mathbf{Y}^{*}$ for any $a \in \mathbb{R}^{+}$
- allows interchangeability of compositions so that $\Delta\left(\mathbf{Y}, \mathbf{Y}^{*}\right) = \Delta\left(\mathbf{Y}^{*}, \mathbf{Y}\right)$
- is scale invariant so that $\Delta\left(a\mathbf{Y}, a^{*}\mathbf{Y}^{*}\right) = \Delta\left(\mathbf{Y}, \mathbf{Y}^{*}\right)$ for every $a, a^{*} \in \mathbb{R}^{+}$
- is perturbation invariant so that $\Delta\left(\mathbf{p} \oplus \mathbf{Y}, \mathbf{p} \oplus \mathbf{Y}^{*}\right) = \Delta\left(\mathbf{Y}, \mathbf{Y}^{*}\right)$ for every perturbation $\mathbf{p}$ (the perturbation operation has been defined in Definition 2.1.1)
- is permutation invariant so that $\Delta\left(\mathbb{P}\mathbf{Y}, \mathbb{P}\mathbf{Y}^{*}\right) = \Delta\left(\mathbf{Y}, \mathbf{Y}^{*}\right)$ for every matrix of permutations $\mathbb{P}$
- satisfies subcompositional dominance so that $\Delta\left(\mathbf{Y}, \mathbf{Y}^{*}\right) \geq \Delta\left(\mathbf{Y}^{s}, \mathbf{Y}^{*s}\right)$, where $\mathbf{Y}^{s}$

and $\mathbf{Y}^{*s}$ are subcompositions of $\mathbf{Y}$ and $\mathbf{Y}^*$ respectively. A subcomposition is a 'composition in a simplex of lower dimension than that of the full composition' (Aitchison, 1986, p. 34).

In considering the goodness of fit of a model, the distance of interest is that between the matrix of compositional data $\mathbb{Y}$ and its corresponding fitted values. Let the fitted values be denoted by $\widehat{\mathbb{P}}^*$. Aitchison's distance measure is then given by

$$\Delta\left(\mathbb{Y}, \widehat{\mathbb{P}}^*\right) = \left[\sum_{j''=1}^{J} \underbrace{\sum_{j=1}^{J}}_{j<j''} \sum_{i=1}^{n} \left[\log\left(\frac{Y_{ij}}{Y_{ij''}}\right) - \log\left(\frac{\hat{p}^*_{ij}}{\hat{p}^*_{ij''}}\right)\right]^2\right]^{\frac{1}{2}}. \tag{3.29}$$

Now, (3.29) may alternatively be rewritten as

$$\begin{aligned}
\Delta\left(\mathbb{Y}, \widehat{\mathbb{P}}^*\right) &= \left[\sum_{j''=1}^{J} \underbrace{\sum_{j=1}^{J}}_{j<j''} \sum_{i=1}^{n} \left[\log\left(\frac{Y_{ij}}{\hat{p}^*_{ij}}\right) - \log\left(\frac{Y_{ij''}}{\hat{p}^*_{ij''}}\right)\right]^2\right]^{\frac{1}{2}} \\
&= \left[J \sum_{j=1}^{J} \sum_{i=1}^{n} \left[\log\left(\frac{Y_{ij}}{\hat{p}^*_{ij}}\right) - \frac{1}{J}\sum_{j'=1}^{J} \log\left(\frac{Y_{ij'}}{\hat{p}^*_{ij'}}\right)\right]^2\right]^{\frac{1}{2}} \\
&= \left[J \sum_{j=1}^{J} \sum_{i=1}^{n} R^{*2}_{ij}\right]^{\frac{1}{2}},
\end{aligned} \tag{3.30}$$

showing that the centred residuals R^*_{ij} may indeed be used to provide a measure of fit of a model that has been fitted to compositional data. The proof showing that the first two lines in (3.30) are equal is presented in Appendix C.

3.6.2 Residuals and Distance Measure based on the Generalized Wedderburn Method

When the generalized Wedderburn method is used, in (2.73) we have seen that the Pearson residual takes the form

$$R_{ij} = \frac{Y_{ij}}{\hat{p}_{ij}} - \frac{1}{J}\sum_{j'=1}^{J} \frac{Y_{ij'}}{\hat{p}_{ij'}}. \tag{3.31}$$

Analogous to the centred residuals R^*_{ij}, from estimating equations (2.55) it may be noticed that the Pearson residuals (3.31) are also the working residuals and that, for all j and s,

they also satisfy

$$\sum_{i=1}^{n} R_{ij} x_{is} = 0, \tag{3.32}$$

meaning that the Pearson residuals are also orthogonal to the columns of the design matrix $\mathbb{X}$.

The fact that under the generalized Wedderburn approach, the Pearson residuals are equivalent to the working residuals is appealing as this is a generalization of the equivalence of the Pearson and working residuals for the model used by Wedderburn (1974) for $J = 2$. Another correspondence involving the Pearson residuals and the model used by Wedderburn (1974) may be obtained by considering the Pearson chi-square statistic. Focusing on the special case $J = 2$, let $\mathbf{R}_i$ denote the vector of Pearson residuals for case i. Then, for $J = 2$, $\mathbf{R}_i$ is given by

$$\mathbf{R}_i = \frac{1}{2} \frac{Y_{i1} - \hat{p}_{i1}}{\hat{p}_{i1}\left(1 - \hat{p}_{i1}\right)} \begin{pmatrix} 1 \\ -1 \end{pmatrix}.$$

For $J = 2$, the Pearson chi-square statistic is thus given by

$$\frac{1}{4} \sum_{i=1}^{n} \left[\frac{\left(Y_{i1} - \hat{p}_{i1}\right)^2}{\hat{p}_{i1}^2 \left(1 - \hat{p}_{i1}\right)^2} + \frac{\left(Y_{i1} - \hat{p}_{i1}\right)^2}{\hat{p}_{i1}^2 \left(1 - \hat{p}_{i1}\right)^2} \right] = \frac{1}{2} \sum_{i=1}^{n} \frac{\left(Y_{i1} - \hat{p}_{i1}\right)^2}{\hat{p}_{i1}^2 \left(1 - \hat{p}_{i1}\right)^2},$$

which is half the value of the Pearson chi-square statistic obtained by Wedderburn (1974). To understand why the Pearson chi-square statistic obtained using the generalized Wedderburn method is half that obtained by Wedderburn (1974), we turn our attention to the asymptotic variance-covariance matrices of $\widehat{\boldsymbol{\gamma}}$ obtained for $J = 2$ under the generalized Wedderburn method and the model used by Wedderburn (1974). The asymptotic variance-covariance matrix obtained using the model in Wedderburn (1974) is given by

$$\text{Var}\left(\widehat{\boldsymbol{\gamma}}\right) = \phi_W \left(\mathbb{X}'\mathbb{X}\right)^{-1}. \tag{3.33}$$

For a J-part composition, if the true variance-covariance matrix of $\mathbf{Y}_i$ is taken to be $\phi \mathbb{V}_{\mathbf{p}_i, \boldsymbol{\Omega}, \mathbb{W}}$, the asymptotic variance-covariance matrix obtained under the generalized Wedderburn method is given by $\mathbb{F}\boldsymbol{\Sigma}\mathbb{F}' \otimes \left(\mathbb{X}'\mathbb{X}\right)^{-1}$. This result will be derived in Section 3.7.2.2. For $J = 2$ and under the assumption that the true variance-covariance matrix of $\mathbf{Y}_i$ is given by $\phi \mathbb{V}_{\mathbf{p}_i, \mathbb{I}_J, \mathbb{I}_J}$, the asymptotic variance-covariance matrix for $\widehat{\boldsymbol{\gamma}}$ reduces to

$$\text{Var}\left(\widehat{\boldsymbol{\gamma}}\right) = 2\phi \left(\mathbb{X}'\mathbb{X}\right)^{-1}. \tag{3.34}$$

On comparing (3.34) with (3.33) it may thus be noticed that the asymptotic variance-covariance matrices achieved under the generalized Wedderburn method and the method used in Wedderburn (1974) are of the same form but they have different dispersion param-

eters. The dispersion parameter used in the generalized Wedderburn method is half the dispersion parameter used in Wedderburn (1974). Now, in the quasi-likelihood estimation framework, the dispersion parameter is typically estimated using the Pearson chi-square statistic. Since the dispersion parameters that correspond to the two methods differ by a factor of two, the corresponding Pearson chi-square statistics will also differ by a factor of two, explaining why the Pearson chi-square statistic obtained using the generalized Wedderburn method is half that based on Wedderburn (1974).

In this section, let $\widehat{\mathbb{P}}$ denote the matrix of fitted values. A multivariate version of the Pearson chi-square statistic may be computed by squaring and summing the Pearson residuals R_{ij} for all i and j. This leads to the the newly proposed directed distance measure $\Delta\left(\mathbb{Y}, \widehat{\mathbb{P}}\right)$ to be used to check goodness of fit of a model fitted to compositional data and which is based on the generalized Wedderburn method. The measure $\Delta\left(\mathbb{Y}, \widehat{\mathbb{P}}\right)$ is computed as

$$\Delta\left(\mathbb{Y}, \widehat{\mathbb{P}}\right) = \left[J \sum_{i=1}^{n} \sum_{j=1}^{J} R_{ij}^2\right]^{\frac{1}{2}}, \tag{3.35}$$

where the square root has been introduced so that this distance measure has a similar meaning to the distance measure developed by Aitchison (1992).

The distance measure $\Delta\left(\mathbb{Y}, \widehat{\mathbb{P}}\right)$ adheres to all the properties of a distance measure listed by Aitchison (1992), except for the interchangeability of the compositions $\mathbf{Y}_i$ and $\widehat{\mathbf{p}}_i$ and the property of subcompositional dominance. The lack of interchangeability in the distance measure $\Delta\left(\mathbb{Y}, \widehat{\mathbb{P}}\right)$ should not however be viewed as a problem, as to check the goodness of fit of a model it makes sense to check how far is a vector of fitted values $\widehat{\mathbf{p}}_i$ from the vector of compositions $\mathbf{Y}_i$ but not vice versa. The requirement for subcompositional dominance is too strict for use with the generalized Wedderburn approach. Under the generalized Wedderburn approach, the model assumptions that are used to analyze subcompositions are consistent with those used to analyze a full composition. For example, for some reference component J, the logits $\log\left(E\left(Y_{ij}\right)/E\left(Y_{iJ}\right)\right)$ are all modeled as $\mathbf{x}_i^{'}\boldsymbol{\gamma}$ if either a full composition or a subcomposition is analyzed. However, the parameter estimates that are obtained from analyzing a full composition are not in general the same as those obtained when a subcomposition is analyzed. This property might be considered by some to be a shortcoming of the model. We however argue that the value of say Y_{i2} being irrelevant to inference about $E\left(Y_{i1}\right)/E\left(Y_{i3}\right)$ when considering a three-part composition is hard to justify and might even be undesirable in general. Consider the following example in terms of logratios in relation to this argument.

Suppose that, for $i = 1, \ldots, n$, compositions $\left(Y_{i1}, Y_{i2}, Y_{i3}\right)^{'}$ are independently logistic normally distributed with the special mean structure $E\left(\log\left(Y_{i1}/Y_{i3}\right)\right) = E\left(\log\left(Y_{i2}/Y_{i3}\right)\right) = \mu$. This example, despite being quite unrealistic as a model to be used with real life data, shows that optimal estimation of μ combines the sample means of $\log\left(Y_{i1}/Y_{i3}\right)$ and $\log\left(Y_{i2}/Y_{i3}\right)$, thus making use of information from all parts of the composition. The adop-

tion of subcompositional dominance as a general *principle* in the analysis of compositional data is therefore unwarranted.

3.7 Properties of Estimators under the Two Models

Since the generalized Wedderburn method is being proposed as an alternative to Aitchison's approach to model the influence of explanatory variables on compositional response variables, particularly when there are zeros in the data, it is important for us to study the properties of the estimators used in the two approaches. Focus is directed towards the difference in the β parameters $(\boldsymbol{\gamma}_1^{'}, \ldots, \boldsymbol{\gamma}_{J-1}^{'})$, taking the last component as reference component.

3.7.1 Properties of Estimators under Aitchison's Model

Let $\widehat{\boldsymbol{\beta}_j^*}$ be the estimator of $\boldsymbol{\beta}_j^*$. Using (3.13) and (3.14),

$$\log\left(\dot{\mathbf{Y}}_{(j)}\right) = \dot{\boldsymbol{\theta}} + \mathbb{X}\boldsymbol{\beta}_j^* + \mathbf{E}_{(j)}^*$$

where $\dot{\mathbf{Y}}_{(j)} = \left(\dot{Y}_{1j}, \ldots, \dot{Y}_{nj}\right)^{'}$, $\dot{\boldsymbol{\theta}} = \left(\dot{\theta}_1, \ldots, \dot{\theta}_n\right)^{'}$ and $\mathbf{E}_{(j)}^* = (E_{1j}, \ldots, E_{nj})^{'}$. It is assumed that for $\mathbf{E}_i^* = (E_{i1}^*, \ldots, E_{iJ}^*)^{'}$, $E(\mathbf{E}_i^*) = \mathbf{0}$ and $\text{Var}(\mathbf{E}_i^*) = \boldsymbol{\Psi}$. The expectation $E\left(\mathbf{E}_{(j)}^*\right)$ is thus also equal to $\mathbf{0}$, from which it follows that

$$\begin{aligned} E\left(\widehat{\boldsymbol{\beta}_j^*}\right) &= \left(\mathbb{X}'\mathbb{X}\right)^{-1}\mathbb{X}'E\left(\log\left(\dot{\mathbf{Y}}_{(j)}\right)\right) \\ &= \left(\mathbb{X}'\mathbb{X}\right)^{-1}\mathbb{X}'\left[\dot{\boldsymbol{\theta}} + \mathbb{X}\boldsymbol{\beta}_j^*\right] \\ &= \boldsymbol{\beta}_j^* + \left(\mathbb{X}'\mathbb{X}\right)^{-1}\mathbb{X}'\dot{\boldsymbol{\theta}}. \end{aligned}$$

So $\widehat{\boldsymbol{\beta}_j^*}$ is not an unbiased estimator of $\boldsymbol{\beta}_j^*$ but because of the fact that the bias $\left(\mathbb{X}'\mathbb{X}\right)^{-1}\mathbb{X}'\dot{\boldsymbol{\theta}}$ is constant between components j, then

$$\begin{aligned} E\left(\widehat{\boldsymbol{\gamma}_j^*}\right) &= E\left(\widehat{\boldsymbol{\beta}_j^*}\right) - E\left(\widehat{\boldsymbol{\beta}_J^*}\right) \\ &= \boldsymbol{\beta}_j^* - \boldsymbol{\beta}_J^* \\ &= \boldsymbol{\gamma}_j^*, \end{aligned} \tag{3.36}$$

showing that $\widehat{\boldsymbol{\gamma}_j^*}$ is an unbiased estimator of $\boldsymbol{\gamma}_j^*$. Thus, under Aitchison's approach and for $(j = 1, \ldots, J-1)$, $\widehat{\gamma}_{j0}^*, \widehat{\gamma}_{j1}, \ldots, \widehat{\gamma}_{jp}$ are unbiased estimators of $\gamma_{j0}^*, \gamma_{j1}, \ldots, \gamma_{jp}$ respectively.

Focus will next be directed towards obtaining an expression for $\text{Var}(\widehat{\boldsymbol{\gamma}}^*)$, where $\boldsymbol{\gamma}^* = \left(\boldsymbol{\gamma}_1^{*'}, \ldots, \boldsymbol{\gamma}_{J-1}^{*'}\right)^{'}$, $\boldsymbol{\gamma}_j^* = \left(\gamma_{j0}^*, \gamma_{j1}, \ldots, \gamma_{jp}\right)^{'}$ for $j = (1, \ldots, J-1)$ and $\widehat{\boldsymbol{\gamma}}^*$ is the estimator

of $\boldsymbol{\gamma}^*$.

So as to better understand the setup that will be used to obtain the variance-covariance matrix of $\widehat{\boldsymbol{\gamma}}^*$, it is important to appreciate the fact that the estimate $\widehat{\boldsymbol{\gamma}}_j^*$ may be obtained in two ways. It is either obtained by solving separately for $\widehat{\boldsymbol{\beta}}_j^*$ and $\widehat{\boldsymbol{\beta}}_J^*$, and then taking their differences. Otherwise, $\widehat{\boldsymbol{\gamma}}_j^*$ may also be obtained directly, by using linear regression modeling on the adjusted vector of working variates

$$\mathbf{W}_{(j)} = \mathbf{Z}_{(j)}^* - \mathbf{Z}_{(J)}^*, \tag{3.37}$$

where $\mathbf{W}_{(j)} = (W_{1j}, \ldots, W_{nj})^{'}$ and for $i = 1, \ldots, n$,

$$W_{ij} = \log\left(\frac{Y_{ij}}{Y_{iJ}}\right). \tag{3.38}$$

In order to obtain $\text{Var}\,(\widehat{\boldsymbol{\gamma}}^*)$, we will focus on the latter procedure. So consider the $(J-1)$-vector of logratios

$$\mathbf{W}_i = \left(\log\left(\frac{Y_{i1}}{Y_{iJ}}\right), \ldots, \log\left(\frac{Y_{i,J-1}}{Y_{iJ}}\right)\right)^{'} = \mathbb{F}\log\left(\mathbf{Y}_i\right) = \mathbb{F}\log\left(\dot{\mathbf{Y}}_i\right) \tag{3.39}$$

where the variance-covariance matrix of $\log\left(\dot{\mathbf{Y}}_i\right)$ is $\boldsymbol{\Psi}$. On modeling the logratios $\mathbf{W}_i$ directly, we can estimate the vector of parameters $\boldsymbol{\gamma}^*$ by means of the generalized least squares estimator

$$\widehat{\boldsymbol{\gamma}}^* = \left(\left(\mathbb{I}_{J-1} \otimes \mathbb{X}\right)^{'} \left(\text{Var}\left(\mathbf{W}\right)\right)^{-1} \left(\mathbb{I}_{J-1} \otimes \mathbb{X}\right)\right)^{-1} \left(\mathbb{I}_{J-1} \otimes \mathbb{X}\right)^{'} \left(\text{Var}\left(\mathbf{W}\right)\right)^{-1} \mathbf{W}$$

where $\mathbf{W} = \left(\mathbf{W}_{(1)}^{'}, \ldots, \mathbf{W}_{(J-1)}^{'}\right)^{'}$, $\mathbf{W}_{(j)} = (W_{1j}, \ldots, W_{nj})^{'}$ for $(j = 1, \ldots, J-1)$ and

$$\begin{aligned}
\text{Var}\left(\mathbf{W}\right) &= \text{Var}\left(\mathbf{W}_i\right) \otimes \mathbb{I}_n \\
&= \text{Var}\left(\mathbb{F}\log\left(\dot{\mathbf{Y}}_i\right)\right) \otimes \mathbb{I}_n \\
&= \mathbb{F}\,\text{Var}\left(\log\left(\dot{\mathbf{Y}}_i\right)\right)\mathbb{F}^{'} \otimes \mathbb{I}_n \\
&= \mathbb{F}\boldsymbol{\Psi}\mathbb{F}^{'} \otimes \mathbb{I}_n.
\end{aligned}$$

Thus

$$\begin{aligned}
\widehat{\boldsymbol{\gamma}}^* &= \left(\left(\mathbb{I}_{J-1} \otimes \mathbb{X}\right)^{'} \left(\left(\mathbb{F}\boldsymbol{\Psi}\mathbb{F}^{'}\right)^{-1} \otimes \mathbb{I}_n\right) \left(\mathbb{I}_{J-1} \otimes \mathbb{X}\right)\right)^{-1} \left(\mathbb{I}_{J-1} \otimes \mathbb{X}\right)^{'} \left(\left(\mathbb{F}\boldsymbol{\Psi}\mathbb{F}^{'}\right)^{-1} \otimes \mathbb{I}_n\right) \mathbf{W} \\
&= \left[\mathbb{I}_{J-1} \otimes \left(\mathbb{X}^{'}\mathbb{X}\right)^{-1}\mathbb{X}^{'}\right] \mathbf{W}
\end{aligned} \tag{3.40}$$

and the variance-covariance matrix of $\widehat{\boldsymbol{\gamma}}^*$ is given by

$$\begin{aligned}
\text{Var}\left(\widehat{\boldsymbol{\gamma}}^*\right) &= \left(\mathbb{I}_{J-1} \otimes \left(\mathbb{X}'\mathbb{X}\right)^{-1}\mathbb{X}'\right) \text{Var}\left(\mathbf{W}\right) \left(\mathbb{I}_{J-1} \otimes \left(\mathbb{X}'\mathbb{X}\right)^{-1}\mathbb{X}'\right)' \\
&= \left(\mathbb{I}_{J-1} \otimes \left(\mathbb{X}'\mathbb{X}\right)^{-1}\mathbb{X}'\right) \left(\mathbb{F}\boldsymbol{\Psi}\mathbb{F}' \otimes \mathbb{I}_n\right) \left(\mathbb{I}_{J-1} \otimes \left(\mathbb{X}'\mathbb{X}\right)^{-1}\mathbb{X}'\right)' \\
&= \mathbb{F}\boldsymbol{\Psi}\mathbb{F}' \otimes \left(\mathbb{X}'\mathbb{X}\right)^{-1}. \qquad (3.41)
\end{aligned}$$

Now, despite stating that $\boldsymbol{\gamma}^*$ is estimated using the GLS estimator (3.40), it should be noted that the estimator $\widehat{\boldsymbol{\gamma}}^*$ is actually free of the variance-covariance matrix $\text{Var}\left(\mathbf{W}\right)$. The presented GLS estimator may thus also be called an OLS estimator. The variance-covariance matrix $\text{Var}\left(\widehat{\boldsymbol{\gamma}}^*\right)$ is however not free of $\text{Var}\left(\mathbf{W}\right)$. The invariance of the estimator $\widehat{\boldsymbol{\gamma}}^*$ to the variance-covariance $\text{Var}\left(\mathbf{W}\right)$ under Aitchison's approach, is directly analogous to the invariance of the estimator $\widehat{\boldsymbol{\gamma}}$ to the working variance-covariance structure used under the generalized Wedderburn approach (refer to Section 2.4.4). The invariance property of these estimators is analogous to the well-established invariance property of GLS estimators in multivariate linear regression (e.g. Mardia et al., 1979, p. 173).

It is also important to mention that since $\widehat{\boldsymbol{\gamma}}^*$ is a GLS estimator, $\widehat{\boldsymbol{\gamma}}^*$ is the best (in terms of the variance) linear unbiased estimator in the class of all linear and unbiased estimators. The estimator $\widehat{\boldsymbol{\gamma}}^*$ is also consistent and asymptotically multivariate normal with mean vector $\boldsymbol{\gamma}^*$ and variance-covariance matrix $\mathbb{F}\boldsymbol{\Psi}\mathbb{F}' \otimes \left(\mathbb{X}'\mathbb{X}\right)^{-1}$. If the error vector $\mathbf{E}_i^*$ is multivariate normally distributed, $\widehat{\boldsymbol{\gamma}}^*$ is multivariate normally distributed with mean vector $\boldsymbol{\gamma}^*$ and variance-covariance matrix $\mathbb{F}\boldsymbol{\Psi}\mathbb{F}' \otimes \left(\mathbb{X}'\mathbb{X}\right)^{-1}$. If the multivariate normality assumption of the error vectors holds, $\widehat{\boldsymbol{\gamma}}^*$ is also fully efficient.

3.7.2 Properties of Estimators under the Generalized Wedderburn Model

3.7.2.1 General Properties of the Estimators

As has been mentioned at the end of Section 2.7, since $\widehat{\boldsymbol{\gamma}}$ is a GEE estimator, it inherits the properties of GEE estimators as in Liang and Zeger (1986). The estimator $\widehat{\boldsymbol{\gamma}}$ is thus asymptotically unbiased with a bias of order $O\left(\frac{1}{n}\right)$ (McCullagh, 1983). It is optimal (has the smallest generalized variance) in the class of linear unbiased estimating equations (McCullagh, 1983). It is also consistent and asymptotically multivariate normal with mean vector $\boldsymbol{\gamma}$ and variance-covariance matrix $\mathbb{F}\boldsymbol{\Sigma}\mathbb{F}' \otimes \left(\mathbb{X}'\mathbb{X}\right)^{-1}$, where $\boldsymbol{\Sigma}$ is as defined in equation (2.62). The derivation to show that the asymptotic variance-covariance matrix $\text{Var}\left(\widehat{\boldsymbol{\gamma}}\right) = \mathbb{F}\boldsymbol{\Sigma}\mathbb{F}' \otimes \left(\mathbb{X}'\mathbb{X}\right)^{-1}$ is given in the following section.

3.7.2.2 Derivation of the Model-Based Asymptotic Variance-Covariance Matrix $\text{Var}(\widehat{\boldsymbol{\gamma}})$

From Section 2.7, $\phi\mathbb{V}_{\mathbf{p}_i,\boldsymbol{\Omega},\mathbb{W}} = \left(\mathbb{P}_i - \mathbf{p}_i\mathbf{p}_i^{'}\right)\boldsymbol{\Sigma}\left(\mathbb{P}_i - \mathbf{p}_i\mathbf{p}_i^{'}\right)$, where $\phi\mathbb{V}_{\mathbf{p}_i,\boldsymbol{\Omega},\mathbb{W}}$ is the working variance-covariance structure assumed under the generalized Wedderburn model. If $\phi\mathbb{V}_{\mathbf{p}_i,\boldsymbol{\Omega},\mathbb{W}}$ is assumed to be the true variance-covariance matrix of $\mathbf{Y}_i$, then

$$\text{Var}(\widehat{\boldsymbol{\gamma}}) = \phi\left(\sum_{i=1}^{n}\mathbb{D}_i^{'}\mathbb{V}_{\mathbf{p}_i,\boldsymbol{\Omega},\mathbb{W}}^{-}\mathbb{D}_i\right)^{-1}, \tag{3.42}$$

where $\mathbb{D}_i$ is as defined on Pg 42.

In order to avoid dealing with singular matrices in the derivation, $\text{Var}(\widehat{\boldsymbol{\gamma}})$ will be reexpressed as

$$\text{Var}(\widehat{\boldsymbol{\gamma}}) = \phi\left(\sum_{i=1}^{n}\left[\mathbb{D}_i\right]_F^{'}\left[\mathbb{V}_{\mathbf{p}_i,\boldsymbol{\Omega},\mathbb{W}}\right]_F^{-1}\left[\mathbb{D}_i\right]_F\right)^{-1},$$

where

$$\left[\mathbb{D}_i\right]_F = \mathbb{F}\left(\mathbb{P}_i - \mathbf{p}_i\mathbf{p}_i^{'}\right)\mathbb{X}_i$$

and

$$\phi\left[\mathbb{V}_{\mathbf{p}_i,\boldsymbol{\Omega},\mathbb{W}}\right]_F = \mathbb{F}\left(\mathbb{P}_i - \mathbf{p}_i\mathbf{p}_i^{'}\right)\boldsymbol{\Sigma}\left(\mathbb{P}_i - \mathbf{p}_i\mathbf{p}_i^{'}\right)\mathbb{F}^{'}.$$

The matrix $\mathbb{F}$ is the $(J-1)\times J$ matrix of contrasts defined in Definition 3.3.1, $\left[\mathbb{D}_i\right]_F$ is a $(J-1)(p+1)\times(J-1)$ matrix which applies contrasts on the matrix of derivatives $\mathbb{D}_i$, and $\left[\mathbb{V}_{\mathbf{p}_i,\boldsymbol{\Omega},\mathbb{W}}\right]_F$ is a $(J-1)\times(J-1)$ matrix which applies contrasts on the variance-covariance matrix of $\mathbf{Y}_i$. The use of contrasts in this situation removes the redundancy attributed to the sum-to-one constraint of the proportions $(p_{i1},\ldots,p_{iJ})$ in the matrix of derivatives and the redundancy attributed to the sum-to-one constraint of the compositional response variables in the variance-covariance matrix of $\mathbf{Y}_i$. By removing the redundancy it is thus possible to derive the expression of $\text{Var}(\widehat{\boldsymbol{\gamma}})$ by using inverses rather than generalized inverses.

Hence

$$\begin{aligned}\text{Var}(\widehat{\boldsymbol{\gamma}}) = &\left(\sum_{i=1}^{n}\mathbb{X}_i^{'}\left(\mathbb{P}_i - \mathbf{p}_i\mathbf{p}_i^{'}\right)\mathbb{F}^{'}\left[\mathbb{F}\left(\mathbb{P}_i - \mathbf{p}_i\mathbf{p}_i^{'}\right)\boldsymbol{\Sigma}\left(\mathbb{P}_i - \mathbf{p}_i\mathbf{p}_i^{'}\right)\mathbb{F}^{'}\right]^{-1}\right.\\ &\left.\times\mathbb{F}\left(\mathbb{P}_i - \mathbf{p}_i\mathbf{p}_i^{'}\right)\mathbb{X}_i\right)^{-1}.\end{aligned} \tag{3.43}$$

Now, the rows of $\mathbb{F}$ form a basis for the subspace of all contrasts in $\mathbb{R}^J$ since the rows of $\mathbb{F}$ sum to 0 and the rows are linearly independent of each other. The matrix $\mathbb{F}\left(\mathbb{P}_i - \mathbf{p}_i\mathbf{p}_i^{'}\right)$ also has rows which form a basis for the same subspace of contrasts in $\mathbb{R}^J$. The latter follows since the rows of $\mathbb{F}\left(\mathbb{P}_i - \mathbf{p}_i\mathbf{p}_i^{'}\right)$ also sum to 0 and are linearly independent. Proofs for the latter two properties may be found in Appendix D.

Since the rows of $\mathbb{F}$ and the rows of $\mathbb{F}\left(\mathbb{P}_i - \mathbf{p}_i\mathbf{p}_i{}'\right)$ each form a basis for the same subspace, then from standard linear algebra it follows that there must exist an invertible $(J-1)\times(J-1)$ matrix, say $\mathbb{G}_i$, such that

$$\mathbb{F}\left(\mathbb{P}_i - \mathbf{p}_i\mathbf{p}_i{}'\right) = \mathbb{G}_i\mathbb{F}. \tag{3.44}$$

Substituting $\mathbb{G}_i\mathbb{F}$ for $\mathbb{F}\left(\mathbb{P}_i - \mathbf{p}_i\mathbf{p}_i{}'\right)$ in (3.43) leads to

$$\begin{aligned}
\text{Var}\left(\widehat{\boldsymbol{\gamma}}\right) &= \left(\sum_{i=1}^{n} \mathbb{X}_i'\mathbb{F}'\mathbb{G}_i'\left[\mathbb{G}_i\mathbb{F}\boldsymbol{\Sigma}\mathbb{F}'\mathbb{G}_i'\right]^{-1}\mathbb{G}_i\mathbb{F}\mathbb{X}_i\right)^{-1} \\
&= \left(\sum_{i=1}^{n} \mathbb{X}_i'\mathbb{F}'\left[\mathbb{F}\boldsymbol{\Sigma}\mathbb{F}'\right]^{-1}\mathbb{F}\mathbb{X}_i\right)^{-1} \\
&= \left(\left[\mathbb{F}\boldsymbol{\Sigma}\mathbb{F}'\right]^{-1}\otimes\mathbb{X}'\mathbb{X}\right)^{-1} \\
&= \mathbb{F}\boldsymbol{\Sigma}\mathbb{F}'\otimes\left(\mathbb{X}'\mathbb{X}\right)^{-1}
\end{aligned} \tag{3.45}$$

where $\mathbb{X}$ is the design matrix.

The final expression in (3.45) shows that the asymptotic variance-covariance matrix of $\widehat{\boldsymbol{\gamma}}$ under the generalized Wedderburn model is in the same form as that obtained using Aitchison's regression model (refer to equation (3.41)).

3.8 Comparison of Asymptotic Efficiency under the Two Models

Aitchison's model and the generalized Wedderburn model may both be used to analyze the influence of explanatory variables on compositional response variables. In Section 3.4, however, it has been shown that the two approaches are based on two different mean models, meaning that they estimate different parameters. Nevertheless, if the response variables do not exhibit any dependency on explanatory variables $X_1,\ldots,X_p$, that would correspond to the corresponding model coefficients being equal to zero in both models simultaneously.

On using Aitchison's approach, if the logratios are multivariate normally distributed, the ordinary least squares estimator, $\widehat{\boldsymbol{\gamma}}^*$, of the model coefficients may equivalently be called the maximum likelihood estimator. Maximum likelihood estimators are well understood and have desirable properties under the model.

Given a correct specification of the mean, the GEE estimator $\widehat{\boldsymbol{\gamma}}$, achieved through the generalized Wedderburn approach is consistent and asymptotically unbiased. A GEE estimator may not however be as efficient as the maximum likelihood estimator used in

Aitchison's approach under the assumption of multivariate normality of the logratios. It is thus of interest to compare how the efficiency of the GEE estimator fares in relation to the maximum likelihood estimator used in Aitchison's approach.

A comparison of efficiency of the estimators is carried out by means of a small simulation study with null effects. The data is generated using Aitchison's model with 3-part compositional response variables and a standard normally distributed continuous random variable as the explanatory variable. The coefficients corresponding to the continuous variable are taken to be zero. The simulation study considers the comparison of efficiency for these coefficients in relation to varying sample sizes, coefficients of variation $\sqrt{\phi\omega_j}, (j = 1, \ldots, J)$ and correlations Corr $\left(\dot{Y}_j, \dot{Y}_{j'}\right) = \alpha_{jj'}, (j \neq j')$.

3.8.1 The Simulation Setup

One hundred thousand samples are generated for each combination of sample size, coefficients of variation and correlation. Three different sample sizes (60, 180 and 600), two different sets of coefficients of variation ($(5\%, 5\%, 20\%)$ and $(30\%, 30\%, 60\%)$) and three different correlations (independence, 0.3 and 0.7) are considered in this simulation study.

Compositional data is generated by considering that if $\log\left(\dot{\mathbf{Y}}_i\right)$ follows a multivariate normal distribution with mean vector $\boldsymbol{\zeta}_i$ and variance covariance matrix $\boldsymbol{\Psi}$, then $\dot{\mathbf{Y}}_i$ follows a multivariate lognormal distribution with parameters $\boldsymbol{\zeta}_i$ and $\boldsymbol{\Psi}$ and taking the closure of $\dot{\mathbf{Y}}_i$ leads to the vector of compositional response variables $\mathbf{Y}_i$. Knowing the values of $m_i\left(\dot{\theta}_i, \boldsymbol{\beta}_j\right)$, coefficients of variation and correlations, the values of the parameters $\boldsymbol{\zeta}_i$ and $\boldsymbol{\Psi}$ are calculated through equations (3.7), (3.8), (3.9) and (3.10).

The coefficients of variation and correlations are taken to be fixed at the values used in the simulation study. The values for $m_i\left(\dot{\theta}_i, \boldsymbol{\beta}_j\right) = \exp\left(\dot{\theta}_i + \boldsymbol{x}_i'\boldsymbol{\beta}_j\right)$ used for the data generation procedure are obtained by fixing a set of β and $\dot{\theta}$ parameters. The β parameters are taken to be $\beta_{10} = 0.14$, $\beta_{20} = 0.02$, $\beta_{30} = 0.04$, $\beta_{11} = 0$, $\beta_{21} = 0$, $\beta_{31} = 0$. The vector $\dot{\boldsymbol{\theta}}$ is generated using the standard normal distribution.

By taking the third component as reference component, the true γ parameters are calculated by taking the difference $\boldsymbol{\beta}_j - \boldsymbol{\beta}_3$, $j \neq 3$, leading to the values shown in Table 3.1, where γ_{11} and γ_{21} are set equal to 0.

	Component 1	**Component 2**
Intercept	0.1	−0.02
x	0	0

Table 3.1: Table of True γ Parameters

Once $\boldsymbol{\zeta}_i$ and $\boldsymbol{\Psi}$ are calculated, the 3-part compositional data is then generated. Since $\dot{\mathbf{Y}}_i$ are taken to follow a multivariate lognormal distribution, the compositional response variables will not contain any zeros.

Once the data generating procedure and the true γ parameter values are set, the simulation study may be carried out. For each generated sample of data, estimates using Aitchison's approach are obtained by fitting the linear model

$$E\left(\log\left(Y_{ij}\right)\right) = \beta_{j0}^{*} + \beta_{j1}x_i \tag{3.46}$$

for each of the three components. Estimates $\widehat{\boldsymbol{\gamma}}_j^{*}$ are obtained by taking the difference $\widehat{\boldsymbol{\beta}}_j^{*} - \widehat{\boldsymbol{\beta}}_3$, $j \neq 3$, and the resulting fitted values are exponentiated and rescaled so that they adhere to the sum-to-1 constraint.

For each generated sample, estimates using the generalized Wedderburn approach are obtained through an iterative process. The linear model

$$E\left(\log\left(m_i\left(\widehat{\theta}_i, \widehat{\boldsymbol{\beta}}_j\right)\right) + \left(\frac{Y_{ij}}{m_i\left(\widehat{\theta}_i, \widehat{\boldsymbol{\beta}}_j\right)} - 1\right)\right) = \beta_{j0} + \beta_{j1}x_i \tag{3.47}$$

is used to obtain the β estimates. The initial value for each θ_i is taken to be 0 and the initial values of $m_i\left(\widehat{\theta}_i, \widehat{\boldsymbol{\beta}}_j\right)$, $(i = 1, \ldots, n, j = 1, \ldots, J)$ are taken to be the fitted values obtained from (3.46). The initial estimates of $\theta_1, \ldots, \theta_n$ are updated once the initial values of $m_i\left(\widehat{\theta}_i, \widehat{\boldsymbol{\beta}}_j\right)$, $(i = 1, \ldots, n, j = 1, \ldots, J)$, are obtained. The initial values of $m_i\left(\widehat{\theta}_i, \widehat{\boldsymbol{\beta}}_j\right)$ are hence rescaled and used in the linear model (3.47) to obtain updated values of $m_i\left(\widehat{\theta}_i, \widehat{\boldsymbol{\beta}}_j\right)$. This leads to another update in the estimates of $\theta_1, \ldots, \theta_n$ which is used to once again obtain rescaled values of $m_i\left(\widehat{\theta}_i, \widehat{\boldsymbol{\beta}}_j\right)$ and the procedure is repeated until convergence is achieved. The convergence criterion used in the simulation study is

$$\left|m_i\left(\widehat{\theta}_i^{t+1}, \widehat{\boldsymbol{\beta}}_j^{t+1}\right) - m_i\left(\widehat{\theta}_i^{t}, \widehat{\boldsymbol{\beta}}_j^{t}\right)\right| < \epsilon \tag{3.48}$$

where t denotes the iteration number and ϵ is a predefined level of tolerance. In this study, ϵ is set to be equal to 10^{-8} and for convergence to be achieved, the convergence criterion (3.48) has to be satisfied for all i and j. Having achieved convergence, estimates $\widehat{\boldsymbol{\gamma}}_j$ are obtained by taking the difference $\widehat{\boldsymbol{\beta}}_j - \widehat{\boldsymbol{\beta}}_3$, $j \neq 3$.

Under the generalized Wedderburn approach, two different estimates of the variance-covariance matrix $\text{Var}\left(\widehat{\boldsymbol{\gamma}}\right)$ are obtained for each sample; the model-based estimator (2.67) with $\hat{\phi}\widehat{\mathbb{V}_{\mathbf{p}_i,\boldsymbol{\Omega},\mathbb{W}}}$ worked out using (2.74) and the robust estimator of Liang and Zeger (1986) described in Section 2.8.3. This is done in order to be able to compare the performance of the two estimators under various sample sizes, coefficients of variation and correlation coefficients. Such a comparison may be entertained by computing the sample variance for each γ parameter using the γ estimates obtained from the generated samples.

Also for every sample and for both the model-based and robust variance estimators, confidence intervals for each of the γ parameters are computed using the estimated standard errors. The estimated standard errors are obtained by taking the square root of the diagonal elements of the model-based and robust estimates of $\text{Var}\left(\widehat{\boldsymbol{\gamma}}\right)$. For every sample

and every parameter, note is taken of the number of times the true parameter values lie within the confidence intervals obtained throughout. At the end of the simulation study, the coverage probability for every parameter is estimated for both the model-based and robust variance estimator. The empirical coverage probability will be compared with the nominal 95% level. This exercise is also carried out to investigate the performance of the two variance estimators. The coverage probabilities that are closest to 95% are achieved by the better performing variance estimator.

Summarization of the Simulation Results

The estimates that are obtained at the end of the simulation are:

- the biases achieved under the two approaches together with their standard error
- the variance of the γ estimates achieved under the two approaches together with the corresponding standard error
- the average of the estimated Var $(\widehat{\boldsymbol{\gamma}})$ using both model-based and robust variance estimators, under the generalized Wedderburn approach, together with their standard error
- coverage probabilities for every non-intercept γ parameter using both model-based and robust variance estimators under the generalized Wedderburn approach.

Since interest lies in the non-intercept parameters, all the results obtained from the simulation study will focus on the coefficients γ_{11} and γ_{21}. To get an idea of the typical simulated datasets that are used in this study, refer to Appendix E. The ternary diagrams presented in Appendix E have been obtained using the first generated sample for each combination of sample size, correlation coefficient and coefficients of variation.

3.8.2 Simulation Results

The results obtained from the simulation study will be presented in this section. We will start by presenting the resulting γ estimates achieved under the two models together with the corresponding estimated biases and their standard error. Scatter plots of Generalized Wedderburn estimates versus Aitchison estimates for γ_{11} and γ_{21} for all the combinations of sample size, correlation and coefficients of variation are presented in Appendix F. The scatter plots show estimates that fall along the line $y = x$ across all conditions. As expected, the points on the scatter plots are more dispersed when a high coefficient of variation is used.

Estimates obtained for the different sample sizes are displayed in Tables 3.2, 3.3 and 3.4. By looking at Table 3.2, it may be noticed that in relation to the estimated biases achieved with the generalized Wedderburn method, the majority of the standard errors achieved are quite large, showing that if any bias is present, it is not detectable for a simulation of size

10^5. The results do however show some significant downward bias when a sample of size 60 is used together with high coefficients of variation $(30\%, 30\%, 60\%)$ and a correlation of 0.3 or 0.7 in conjunction with the generalized Wedderburn method.

Table 3.3 and Table 3.4 show no evidence of bias when samples of size 180 and 600 are used in conjunction with the generalized Wedderburn method. As expected, no bias has been achieved throughout all conditions, when Aitchison's regression model has been used.

From Tables 3.2, 3.3 and 3.4 it is also worth pointing that there is barely any difference between the standard errors achieved under the two different approaches across all sample sizes. Some of the standard errors achieved using the generalized Wedderburn approach are actually slightly smaller. The variances from which these standard errors are computed are presented in Tables 3.5, 3.6 and 3.7. The variances obtained using the generalized Wedderburn method are very similar to those achieved using Aitchison's approach, with the variances achieved using the generalized Wedderburn approach being sometimes slightly smaller. The GEE estimator thus manages to achieve the same or even slightly better efficiency than the ML estimator across all sample sizes and across all conditions considered. This behaviour might seem quite surprising, particularly when a sample as large as 600 is used and knowing that the ML estimator is well renowned for being a uniformly minimum variance unbiased estimator asymptotically. An explanation of why the GEE estimator may actually attain better efficiency than the ML estimator follows.

In Section 3.7.2.2 it has been shown that the asymptotic variance-covariance matrix $\text{Var}\left(\widehat{\boldsymbol{\gamma}}\right)$ is given by $\mathbb{F}\boldsymbol{\Sigma}\mathbb{F}' \otimes \left(\mathbb{X}'\mathbb{X}\right)^{-1}$ where $\mathbb{F}$ is the matrix of contrasts, $\mathbb{X}$ is the design matrix and $\boldsymbol{\Sigma} = \phi\boldsymbol{\Omega}^{\frac{1}{2}}\mathbb{W}\boldsymbol{\Omega}^{\frac{1}{2}}$. The variance-covariance matrix of $\widehat{\boldsymbol{\gamma}}^*$ achieved using Aitchison's approach is given by $\text{Var}\left(\widehat{\boldsymbol{\gamma}}^*\right) = \mathbb{F}\boldsymbol{\Psi}\mathbb{F}' \otimes \left(\mathbb{X}'\mathbb{X}\right)^{-1}$. Due to similarity of form of the two variance-covariance matrices, as a measure of relative efficiency of the GEE estimator with respect to the ML estimator, we can direct our attention towards the matrices $\boldsymbol{\Sigma}$ and $\boldsymbol{\Psi}$.

As explained in Section 3.8.1, in this simulation study the data generation is based on $\boldsymbol{\Psi}$, where $\boldsymbol{\Psi}$ has been chosen such that its diagonal elements are equal to $\log\left(1+\phi\omega_j\right)$ and its off-diagonal elements are equal to $\log\left(1+\phi\sqrt{\omega_j}\sqrt{\omega_{j'}}\alpha_{jj'}\right)$, $j \neq j'$, where $\phi\omega_j$ and $\phi\sqrt{\omega_j}\sqrt{\omega_{j'}}\alpha_{jj'}$ are respectively the diagonal and off-diagonal elements of $\dot{\boldsymbol{\Sigma}}$ and $\dot{\boldsymbol{\Sigma}}$ is as defined in Section 3.2.

If it was the case that $\dot{\boldsymbol{\Sigma}} = \boldsymbol{\Sigma}$, then the comparison of efficiency being undertaken here would be a direct generalization of Firth (1988, eq. (4)). Let us reexpress Firth (1988, eq. (4)) in terms of modeling compositional response variables. If the latent variable $\dot{Y}_{ij}$ is assumed to follow a lognormal distribution with parameters ζ_{ij} and $\psi_j^2 = \log\left(1+\phi\omega_j\right)$, the efficiency of the quasi-likelihood estimator obtained by modeling the data on the original scale using the multiplicative model, versus the maximum likelihood estimator obtained by modeling the data on the log-scale is given by $\log\left(1+\phi\omega_j\right)/\phi\omega_j$, with the quasi-likelihood estimator always losing some of the efficiency in comparison to the maximum likelihood estimator. On generalizing this property to modeling the full set of compositional variables,

Sample Size: 60					
Correlation: Independence		**Aitchison Estimates**		**Generalized Wedderburn**	
Coefficient of Variation (%)	**Parameter** $\times 10^3$	**Bias** $\times 10^3$	**Standard Error** $\times 10^3$	**Bias** $\times 10^3$	**Standard Error** $\times 10^3$
$(5, 5, 20)$	γ_{11}	-0.104	0.083	-0.157	0.082
	γ_{21}	-0.119	0.083	-0.173	0.083
$(30, 30, 60)$	γ_{11}	0.099	0.256	-0.250	0.244
	γ_{21}	-0.140	0.256	-0.439	0.244
Correlation: 0.3		**Aitchison Estimates**		**Generalized Wedderburn**	
Coefficient of Variation (%)	**Parameter**	**Bias** $\times 10^3$	**Standard Error** $\times 10^3$	**Bias** $\times 10^3$	**Standard Error** $\times 10^3$
$(5, 5, 20)$	γ_{11}	0.150	0.077	0.098	0.076
	γ_{21}	0.166	0.077	0.116	0.076
$(30, 30, 60)$	γ_{11}	-0.406	0.219	-0.637	0.211
	γ_{21}	-0.113	0.219	-0.313	0.211
Correlation: 0.7		**Aitchison Estimates**		**Generalized Wedderburn**	
Coefficient of Variation (%)	**Parameter**	**Bias** $\times 10^3$	**Standard Error** $\times 10^3$	**Bias** $\times 10^3$	**Standard Error** $\times 10^3$
$(5, 5, 20)$	γ_{11}	0.052	0.068	0.011	0.068
	γ_{21}	0.047	0.068	0.008	0.068
$(30, 30, 60)$	γ_{11}	-0.188	0.160	-0.375	0.157
	γ_{21}	-0.194	0.161	-0.370	0.157

Table 3.2: Table of Estimated Bias and Standard Error of the Bias achieved under Aitchison and Generalized Wedderburn Approach using a sample of size 60 assuming independence, correlation of 0.3 and correlation of 0.7 and a simulation size of 10^5

Sample Size: 180					
Correlation: Independence		**Aitchison Estimates**		**Generalized Wedderburn**	
Coefficient of Variation (%)	**Parameter**	**Bias** $\times 10^3$	**Standard Error** $\times 10^3$	**Bias** $\times 10^3$	**Standard Error** $\times 10^3$
$(5, 5, 20)$	γ_{11}	-0.005	0.051	-0.010	0.051
	γ_{21}	0.009	0.051	0.003	0.051
$(30, 30, 60)$	γ_{11}	0.086	0.158	0.089	0.150
	γ_{21}	0.073	0.158	0.069	0.150
Correlation: 0.3		**Aitchison Estimates**		**Generalized Wedderburn**	
Coefficient of Variation (%)	**Parameter**	**Bias** $\times 10^3$	**Standard Error** $\times 10^3$	**Bias** $\times 10^3$	**Standard Error** $\times 10^3$
$(5, 5, 20)$	γ_{11}	-0.035	0.047	-0.039	0.047
	γ_{21}	-0.052	0.047	-0.057	0.047
$(30, 30, 60)$	γ_{11}	0.207	0.134	0.182	0.129
	γ_{21}	0.182	0.135	0.169	0.130
Correlation: 0.7		**Aitchison Estimates**		**Generalized Wedderburn**	
Coefficient of Variation (%)	**Parameter**	**Bias** $\times 10^3$	**Standard Error** $\times 10^3$	**Bias** $\times 10^3$	**Standard Error** $\times 10^3$
$(5, 5, 20)$	γ_{11}	-0.008	0.042	-0.011	0.041
	γ_{21}	0.005	0.042	0.002	0.041
$(30, 30, 60)$	γ_{11}	0.207	0.099	0.187	0.097
	γ_{21}	0.116	0.100	0.094	0.097

Table 3.3: Table of Estimated Bias and Standard Error of the Bias achieved under Aitchison and Generalized Wedderburn Approach using a sample of size 180 assuming independence, correlation of 0.3 and correlation of 0.7 and a simulation size of 10^5

Sample Size: 600					
Correlation: Independence		**Aitchison Estimates**		**Generalized Wedderburn**	
Coefficient of Variation (%)	**Parameter**	**Bias** $\times 10^3$	**Standard Error** $\times 10^3$	**Bias** $\times 10^3$	**Standard Error** $\times 10^3$
$(5, 5, 20)$	γ_{11}	0.019	0.026	0.020	0.026
	γ_{21}	0.015	0.026	0.016	0.026
$(30, 30, 60)$	γ_{11}	0.049	0.079	0.062	0.075
	γ_{21}	0.057	0.079	0.070	0.075
Correlation: 0.3		**Aitchison Estimates**		**Generalized Wedderburn**	
Coefficient of Variation (%)	**Parameter**	**Bias** $\times 10^3$	**Standard Error** $\times 10^3$	**Bias** $\times 10^3$	**Standard Error** $\times 10^3$
$(5, 5, 20)$	γ_{11}	0.003	0.024	0.003	0.024
	γ_{21}	0.003	0.024	0.003	0.024
$(30, 30, 60)$	γ_{11}	-0.020	0.068	-0.020	0.065
	γ_{21}	-0.017	0.068	-0.006	0.065
Correlation: 0.7		**Aitchison Estimates**		**Generalized Wedderburn**	
Coefficient of Variation (%)	**Parameter**	**Bias** $\times 10^3$	**Standard Error** $\times 10^3$	**Bias** $\times 10^3$	**Standard Error** $\times 10^3$
$(5, 5, 20)$	γ_{11}	0.002	0.021	0.002	0.021
	γ_{21}	0.001	0.021	0.000	0.021
$(30, 30, 60)$	γ_{11}	0.072	0.050	0.073	0.049
	γ_{21}	0.028	0.050	0.031	0.049

Table 3.4: Table of Estimated Bias and Standard Error of the Bias achieved under Aitchison and Generalized Wedderburn Approach using a sample of size 600 assuming independence, correlation of 0.3 and correlation of 0.7 and a simulation size of 10^5

it might seem that a similar loss of efficiency for the GEE estimator should also be obtained. This does not however hold true for our case. As explained in Section 2.7, the generalized Wedderburn method relies on the specification of a specific working variance-covariance structure for the compositional variables which is in terms of $\boldsymbol{\Sigma}$, and $\boldsymbol{\Sigma}$ and $\dot{\boldsymbol{\Sigma}}$ can only be equal to each other if there is no dispersion in the model, which is very unrealistic.

In this particular simulation study, $E(Y_{ij}) = p_{ij} = p_j$ since the proportions p_{ij}, for component j, are the same for all cases i. The true distribution of $\mathbf{Y}_i$ is thus the same for all i. We can get a good estimate of the true variance-covariance matrix $\text{Var}(\mathbf{Y}_i)$ by finding the average of the variance-covariance matrices obtained from all the generated samples. Then, by using the assumed form of the variance-covariance matrix of $\mathbf{Y}_i$ under the generalized Wedderburn method, that is,

$$\text{Var}(\mathbf{Y}_i) = \left(\mathbb{P}_i - \mathbf{p}_i\mathbf{p}_i{}'\right)\boldsymbol{\Sigma}\left(\mathbb{P}_i - \mathbf{p}_i\mathbf{p}_i{}'\right), \tag{3.49}$$

we can obtain an estimate of $\boldsymbol{\Sigma}$, using

$$\widehat{\boldsymbol{\Sigma}} = \left(\mathbb{P}_i - \mathbf{p}_i\mathbf{p}_i{}'\right)^{+}\widehat{\text{Var}(\mathbf{Y}_i)}\left(\mathbb{P}_i - \mathbf{p}_i\mathbf{p}_i{}'\right)^{+}, \tag{3.50}$$

where $\mathbf{p}_i$ is the vector of known proportions $(0.358, 0.318, 0.324)$ obtained using equation (3.17) with the values of the γ parameters taken to be the true parameter values shown in Table 3.1 and the values in $\mathbf{x}_i$ being those used in the design matrix in the simulation study, $\mathbb{P}_i = \text{diag}(0.358, 0.318, 0.324)$, $\left(\mathbb{P}_i - \mathbf{p}_i\mathbf{p}_i{}'\right)^{+}$ is the Moore-Penrose pseudoinverse of $\mathbb{P}_i - \mathbf{p}_i\mathbf{p}_i{}'$ and $\widehat{\text{Var}(\mathbf{Y}_i)}$ is an estimate of the true variance-covariance matrix of $\mathbf{Y}_i$.

The estimated variance-covariance matrix which resulted from the 10^5 datasets that were generated using samples of size 180 with coefficients of variation $(30\%, 30\%, 60\%)$ and correlation 0.7 is given by

$$\begin{pmatrix} 0.0033 & 0.0000 & -0.0033 \\ 0.0000 & 0.0028 & -0.0028 \\ -0.0033 & -0.0028 & 0.0061 \end{pmatrix}, \tag{3.51}$$

leading to

$$\widehat{\boldsymbol{\Sigma}} = \begin{pmatrix} 0.0255 & 0.0003 & -0.0285 \\ 0.0003 & 0.0278 & -0.0276 \\ -0.0285 & -0.0276 & 0.0585 \end{pmatrix}. \tag{3.52}$$

Using coefficients of variation $(30\%, 30\%, 60\%)$ and correlation 0.7, the matrix $\boldsymbol{\Psi}$ is given by

$$\boldsymbol{\Psi} = \begin{pmatrix} 0.090 & 0.063 & 0.126 \\ 0.063 & 0.090 & 0.126 \\ 0.126 & 0.126 & 0.360 \end{pmatrix}. \tag{3.53}$$

The fact the GEE estimators under the generalized Wedderburn approach may achieve better efficiency than ML estimators using Aitchison's approach may then be noticed by comparing the variances inside $\widehat{\boldsymbol{\Sigma}}$ with the corresponding values in $\boldsymbol{\Psi}$. The variances in $\widehat{\boldsymbol{\Sigma}}$ are clearly smaller than those in $\boldsymbol{\Psi}$.

Besides studying how the estimates obtained under the generalized Wedderburn approach compare to those obtained using Aitchison's approach, it is also of interest to study how the model-based variance estimator and the robust variance estimator compare to each other in estimating the variance of the γ estimates under the generalized Wedderburn approach. As mentioned previously, to be able to do such a comparison, the sample variance for each γ parameter using the γ estimates obtained from the generated samples has to be computed. The model-based, robust and the estimated sample variance obtained under the generalized Wedderburn approach are summarized in Tables 3.8, 3.9 and 3.10. The model-based variance estimates $\widehat{\text{Var}}(\widehat{\gamma})_M$ are closer on average to the sample variances $\widehat{\text{Var}}(\widehat{\gamma})$ than are the corresponding robust estimates $\widehat{\text{Var}}(\widehat{\gamma})_{LZ}$, across all conditions. The model-based and robust variance estimates become closer with an increase in sample size. With regard to the standard errors of the variance estimates, the majority of the standard errors of the robust estimator are higher than those corresponding to the model-based estimator across all sample sizes. Overall, the standard errors of the two variance estimators are in agreement with equation (2.81) which states that the variance of the robust estimator is greater than or equal to the variance of the model-based estimator as the sample size increases. The findings in this simulation study also agree with the simulation results obtained by Pan (2001b). Pan (2001b) shows that the model-based variance estimator is a more efficient estimator even for smaller sample sizes.

The coverage probabilities computed using the model-based and robust variance estimators are presented in Table 3.11. The values shown in this table show the overall superiority of the model-based estimator being proposed in this thesis over the robust estimator. From Tables 3.8, 3.9 and 3.10 it may be seen that the robust variance estimator underestimates the sample variance throughout. Consequently, the coverage probabilities obtained using the robust variance estimator do not manage to reach 95%. The model-based estimator does lose some efficiency with decreasing sample size but even its lowest coverage probability, 94.32%, achieved with samples of size 60, independence and high coefficients of variation, is still very close to 95%. The coverage probabilities of the model-based estimator are in fact either 95% or very close to it across all conditions. The same cannot be said for the robust estimator.

3.9 Analyzing the Arctic Lake Dataset

The Arctic Lake dataset (see Table 3.12, Data 5 from Appendix D of Aitchison (1986)) is one of the most widely used datasets in compositional data literature (e.g. Aitchison, 1986; Tsagris et al., 2011; Maier, 2014). For this reason, in this section we will illustrate how

Sample Size: 60					
Correlation: Independence		**Aitchison Estimates**		**Generalized Wedderburn**	
Coefficient of Variation (%)	**Parameter**	**Variance** $\times 10^3$	**Standard Error** $\times 10^3$	**Variance** $\times 10^3$	**Standard Error** $\times 10^3$
$(5, 5, 20)$	γ_{11}	0.6886	0.0031	0.6797	0.0030
	γ_{21}	0.6895	0.0031	0.6806	0.0030
$(30, 30, 60)$	γ_{11}	6.5684	0.0294	5.9599	0.0267
	γ_{21}	6.5359	0.0292	5.9601	0.0267
Correlation: 0.3		**Aitchison Estimates**		**Generalized Wedderburn**	
Coefficient of Variation (%)	**Parameter**	**Variance** $\times 10^3$	**Standard Error** $\times 10^3$	**Variance** $\times 10^3$	**Standard Error** $\times 10^3$
$(5, 5, 20)$	γ_{11}	0.5863	0.0026	0.5794	0.0026
	γ_{21}	0.5872	0.0026	0.5804	0.0026
$(30, 30, 60)$	γ_{11}	4.7905	0.0214	4.4521	0.0199
	γ_{21}	4.7848	0.0214	4.4615	0.0200
Correlation: 0.7		**Aitchison Estimates**		**Generalized Wedderburn**	
Coefficient of Variation (%)	**Parameter**	**Variance** $\times 10^3$	**Standard Error** $\times 10^3$	**Variance** $\times 10^3$	**Standard Error** $\times 10^3$
$(5, 5, 20)$	γ_{11}	0.4596	0.0021	0.4557	0.0020
	γ_{21}	0.4608	0.0021	0.4568	0.0020
$(30, 30, 60)$	γ_{11}	2.5616	0.0115	2.4548	0.0110
	γ_{21}	2.5769	0.0115	2.4756	0.0111

Table 3.5: Table of Variance Estimates together with their standard errors achieved under Aitchison and Generalized Wedderburn Approach using a sample of size 60 under independence, correlation of 0.3 and correlation of 0.7 and a simulation of size 10^5

Sample Size: 180					
Correlation: Independence		**Aitchison Estimates**		**Generalized Wedderburn**	
Coefficient of Variation (%)	**Parameter**	**Variance** $\times 10^3$	**Standard Error** $\times 10^3$	**Variance** $\times 10^3$	**Standard Error** $\times 10^3$
$(5, 5, 20)$	γ_{11}	0.26282	0.00118	0.25912	0.00116
	γ_{21}	0.26232	0.00117	0.25866	0.00116
$(30, 30, 60)$	γ_{11}	2.48362	0.01111	2.23537	0.01000
	γ_{21}	2.48219	0.01110	2.24595	0.01004
Correlation: 0.3		**Aitchison Estimates**		**Generalized Wedderburn**	
Coefficient of Variation (%)	**Parameter**	**Variance** $\times 10^3$	**Standard Error** $\times 10^3$	**Variance** $\times 10^3$	**Standard Error** $\times 10^3$
$(5, 5, 20)$	γ_{11}	0.22471	0.00100	0.22199	0.00099
	γ_{21}	0.22501	0.00101	0.22237	0.00099
$(30, 30, 60)$	γ_{11}	1.80392	0.00807	1.66956	0.00747
	γ_{21}	1.81419	0.00811	1.68423	0.00753
Correlation: 0.7		**Aitchison Estimates**		**Generalized Wedderburn**	
Coefficient of Variation (%)	**Parameter**	**Variance** $\times 10^3$	**Standard Error** $\times 10^3$	**Variance** $\times 10^3$	**Standard Error** $\times 10^3$
$(5, 5, 20)$	γ_{11}	0.17336	0.00078	0.17170	0.00077
	γ_{21}	0.17354	0.00078	0.17187	0.00077
$(30, 30, 60)$	γ_{11}	0.98615	0.00441	0.94335	0.00422
	γ_{21}	0.99170	0.00444	0.94990	0.00425

Table 3.6: Table of Variance Estimates together with their standard errors achieved under Aitchison and Generalized Wedderburn Approach using a sample of size 180 under independence, correlation of 0.3 and correlation of 0.7 and a simulation of size 10^5

Sample Size: 600					
Correlation: Independence		**Aitchison Estimates**		**Generalized Wedderburn**	
Coefficient of Variation (%)	**Parameter**	**Variance** $\times 10^3$	**Standard Error** $\times 10^3$	**Variance** $\times 10^3$	**Standard Error** $\times 10^3$
$(5, 5, 20)$	γ_{11}	0.06682	0.00030	0.06589	0.00029
	γ_{21}	0.06714	0.00030	0.06622	0.00030
$(30, 30, 60)$	γ_{11}	0.62314	0.00279	0.56084	0.00251
	γ_{21}	0.62589	0.00280	0.56574	0.00253
Correlation: 0.3		**Aitchison Estimates**		**Generalized Wedderburn**	
Coefficient of Variation (%)	**Parameter**	**Variance** $\times 10^3$	**Standard Error** $\times 10^3$	**Variance** $\times 10^3$	**Standard Error** $\times 10^3$
$(5, 5, 20)$	γ_{11}	0.05721	0.00026	0.05652	0.00025
	γ_{21}	0.05724	0.00026	0.05656	0.00025
$(30, 30, 60)$	γ_{11}	0.45682	0.00204	0.42238	0.00189
	γ_{21}	0.45983	0.00206	0.42657	0.00191
Correlation: 0.7		**Aitchison Estimates**		**Generalized Wedderburn**	
Coefficient of Variation (%)	**Parameter**	**Variance** $\times 10^3$	**Standard Error** $\times 10^3$	**Variance** $\times 10^3$	**Standard Error** $\times 10^3$
$(5, 5, 20)$	γ_{11}	0.04424	0.00020	0.04380	0.00020
	γ_{21}	0.04419	0.00020	0.04375	0.00020
$(30, 30, 60)$	γ_{11}	0.24692	0.00110	0.23622	0.00106
	γ_{21}	0.24566	0.00110	0.23530	0.00105

Table 3.7: Table of Variance Estimates together with their standard errors achieved under Aitchison and Generalized Wedderburn Approach using a sample of size 600 under independence, correlation of 0.3 and correlation of 0.7 and a simulation of size 10^5

Sample Size: 60							
Correlation: Independence				**Model-Based**		**Robust**	
Coefficient of Variation (%)	**Parameter**	$\widehat{\text{Var}}(\widehat{\gamma})$ $\times 10^3$	**Standard Error** $\times 10^3$	**Mean of** $\widehat{\text{Var}}(\widehat{\gamma})_M$ $\times 10^3$	**Standard Error** $\times 10^3$	**Mean of** $\widehat{\text{Var}}(\widehat{\gamma})_{LZ}$ $\times 10^3$	**Standard Error** $\times 10^3$
$(5, 5, 20)$	γ_{11}	0.6886	0.0031	0.67936	0.00040	0.62379	0.00071
	γ_{21}	0.6895	0.0031	0.67931	0.00040	0.62363	0.00071
$(30, 30, 60)$	γ_{11}	6.5684	0.0294	5.83725	0.00317	5.35102	0.00591
	γ_{21}	6.5359	0.0292	5.86261	0.00321	5.36987	0.00593
Correlation: 0.3				**Model-Based**		**Robust**	
Coefficient of Variation (%)	**Parameter**	$\widehat{\text{Var}}(\widehat{\gamma})$ $\times 10^3$	**Standard Error** $\times 10^3$	**Mean of** $\widehat{\text{Var}}(\widehat{\gamma})_M$ $\times 10^3$	**Standard Error** $\times 10^3$	**Mean of** $\widehat{\text{Var}}(\widehat{\gamma})_{LZ}$ $\times 10^3$	**Standard Error** $\times 10^3$
$(5, 5, 20)$	γ_{11}	0.5863	0.0026	0.58299	0.00034	0.53510	0.00061
	γ_{21}	0.58718	0.0026	0.58322	0.00034	0.53504	0.00061
$(30, 30, 60)$	γ_{11}	4.7905	0.0214	4.39332	0.00244	4.03427	0.00450
	γ_{21}	4.7848	0.0214	4.41131	0.00246	4.03998	0.00453
Correlation: 0.7				**Model-Based**		**Robust**	
Coefficient of Variation (%)	**Parameter**	$\widehat{\text{Var}}(\widehat{\gamma})$ $\times 10^3$	**Standard Error** $\times 10^3$	**Mean of** $\widehat{\text{Var}}(\widehat{\gamma})_M$ $\times 10^3$	**Standard Error** $\times 10^3$	**Mean of** $\widehat{\text{Var}}(\widehat{\gamma})_{LZ}$ $\times 10^3$	**Standard Error** $\times 10^3$
$(5, 5, 20)$	γ_{11}	0.4596	0.0021	0.45405	0.00027	0.41657	0.00047
	γ_{21}	0.4608	0.0021	0.45407	0.00027	0.41666	0.00048
$(30, 30, 60)$	γ_{11}	2.5616	0.0115	2.46329	0.00141	2.26197	0.00255
	γ_{21}	2.5769	0.0115	2.46693	0.00141	2.26119	0.00256

Table 3.8: Table of Variance Estimates together with their standard errors achieved under the Generalized Wedderburn Approach using a sample of size 60 assuming independence, correlation of 0.3 and correlation of 0.7 and a simulation of size 10^5

Sample Size: 180							
Correlation: Independence				**Model-Based**		**Robust**	
Coefficient of Variation (%)	**Parameter**	$\widehat{\text{Var}}(\widehat{\gamma})$ $\times 10^3$	**Standard Error** $\times 10^3$	**Mean of** $\widehat{\text{Var}}(\widehat{\gamma})_M$ $\times 10^3$	**Standard Error** $\times 10^3$	**Mean of** $\widehat{\text{Var}}(\widehat{\gamma})_{LZ}$ $\times 10^3$	**Standard Error** $\times 10^3$
$(5, 5, 20)$	γ_{11}	0.26282	0.00118	0.259544	0.000086	0.253775	0.000146
	γ_{21}	0.26232	0.00117	0.259531	0.000086	0.253747	0.000146
$(30, 30, 60)$	γ_{11}	2.48362	0.01111	2.231265	0.000693	2.181171	0.001222
	γ_{21}	2.48219	0.01110	2.243522	0.000701	2.192812	0.001236
Correlation: 0.3				**Model-Based**		**Robust**	
Coefficient of Variation (%)	**Parameter**	$\widehat{\text{Var}}(\widehat{\gamma})$ $\times 10^3$	**Standard Error** $\times 10^3$	**Mean of** $\widehat{\text{Var}}(\widehat{\gamma})_M$ $\times 10^3$	**Standard Error** $\times 10^3$	**Mean of** $\widehat{\text{Var}}(\widehat{\gamma})_{LZ}$ $\times 10^3$	**Standard Error** $\times 10^3$
$(5, 5, 20)$	γ_{11}	0.22471	0.00100	0.222642	0.000074	0.217723	0.000126
	γ_{21}	0.22501	0.00101	0.222646	0.000074	0.217671	0.000126
$(30, 30, 60)$	γ_{11}	1.80392	0.00807	1.679428	0.000533	1.640365	0.000933
	γ_{21}	1.81419	0.00811	1.686327	0.000538	1.647704	0.000938
Correlation: 0.7				**Model-Based**		**Robust**	
Coefficient of Variation (%)	**Parameter**	$\widehat{\text{Var}}(\widehat{\gamma})$ $\times 10^3$	**Standard Error** $\times 10^3$	**Mean of** $\widehat{\text{Var}}(\widehat{\gamma})_M$ $\times 10^3$	**Standard Error** $\times 10^3$	**Mean of** $\widehat{\text{Var}}(\widehat{\gamma})_{LZ}$ $\times 10^3$	**Standard Error** $\times 10^3$
$(5, 5, 20)$	γ_{11}	0.17336	0.00078	0.173634	0.000058	0.169570	0.000098
	γ_{21}	0.17354	0.00078	0.173617	0.000058	0.169599	0.000098
$(30, 30, 60)$	γ_{11}	0.98615	0.00441	0.942115	0.000308	0.921060	0.000530
	γ_{21}	0.99170	0.00444	0.943603	0.000308	0.922013	0.000531

Table 3.9: Table of Variance Estimates together with their standard errors achieved under the Generalized Wedderburn Approach using a sample of size 180 assuming independence, correlation of 0.3 and correlation of 0.7 and a simulation of size 10^5

Sample Size: 600							
Correlation: Independence				**Model-Based**		**Robust**	
Coefficient of Variation (%)	**Parameter**	$\widehat{\text{Var}}(\widehat{\gamma})$ $\times 10^3$	**Standard Error** $\times 10^3$	**Mean of** $\widehat{\text{Var}}(\widehat{\gamma})_M$ $\times 10^3$	**Standard Error** $\times 10^3$	**Mean of** $\widehat{\text{Var}}(\widehat{\gamma})_{LZ}$ $\times 10^3$	**Standard Error** $\times 10^3$
$(5, 5, 20)$	γ_{11}	0.06682	0.00030	0.065477	0.000012	0.065088	0.000020
	γ_{21}	0.06714	0.00030	0.065498	0.000012	0.065113	0.000020
$(30, 30, 60)$	γ_{11}	0.62314	0.00279	0.562979	0.000095	0.559425	0.000166
	γ_{21}	0.62589	0.00280	0.566027	0.000096	0.562387	0.000167
Correlation: 0.3				**Model-Based**		**Robust**	
Coefficient of Variation (%)	**Parameter**	$\widehat{\text{Var}}(\widehat{\gamma})$ $\times 10^3$	**Standard Error** $\times 10^3$	**Mean of** $\widehat{\text{Var}}(\widehat{\gamma})_M$ $\times 10^3$	**Standard Error** $\times 10^3$	**Mean of** $\widehat{\text{Var}}(\widehat{\gamma})_{LZ}$ $\times 10^3$	**Standard Error** $\times 10^3$
$(5, 5, 20)$	γ_{11}	0.05721	0.00026	0.056200	0.000010	0.055844	0.000017
	γ_{21}	0.05724	0.00026	0.056195	0.000010	0.055838	0.000017
$(30, 30, 60)$	γ_{11}	0.45682	0.00204	0.423935	0.000074	0.421174	0.000126
	γ_{21}	0.45983	0.00206	0.425572	0.000074	0.422795	0.000128
Correlation: 0.7				**Model-Based**		**Robust**	
Coefficient of Variation (%)	**Parameter**	$\widehat{\text{Var}}(\widehat{\gamma})$ $\times 10^3$	**Standard Error** $\times 10^3$	**Mean of** $\widehat{\text{Var}}(\widehat{\gamma})_M$ $\times 10^3$	**Standard Error** $\times 10^3$	**Mean of** $\widehat{\text{Var}}(\widehat{\gamma})_{LZ}$ $\times 10^3$	**Standard Error** $\times 10^3$
$(5, 5, 20)$	γ_{11}	0.04424	0.00020	0.043774	0.000008	0.043483	0.000013
	γ_{21}	0.04419	0.00020	0.043775	0.000008	0.043487	0.000013
$(30, 30, 60)$	γ_{11}	0.24692	0.00110	0.237870	0.000042	0.236394	0.000072
	γ_{21}	0.24566	0.00110	0.238291	0.000043	0.236768	0.000072

Table 3.10: Table of Variance Estimates together with their standard errors achieved under the Generalized Wedderburn Approach using a sample of size 600 assuming independence, correlation of 0.3 and correlation of 0.7 and a simulation of size 10^5

Sample Size: 60		**Coverage Probability**					
Coefficient of Variation (%)	**Parameter**	**Correlation Independence**		**Correlation 0.3**		**Correlation 0.7**	
		Model Based	**Robust**	**Model Based**	**Robust**	**Model Based**	**Robust**
$(5, 5, 20)$	γ_{11}	94.52	92.12	94.56	92.21	94.49	92.19
	γ_{21}	94.55	92.15	94.57	92.22	94.44	92.15
$(30, 30, 60)$	γ_{11}	94.33	91.84	94.37	92.04	94.52	92.28
	γ_{21}	94.32	91.84	94.38	92.06	94.51	92.26
Sample Size: 180		**Coverage Probability**					
Coefficient of Variation (%)	**Parameter**	**Correlation Independence**		**Correlation 0.3**		**Correlation 0.7**	
		Model Based	**Robust**	**Model Based**	**Robust**	**Model Based**	**Robust**
$(5, 5, 20)$	γ_{11}	94.76	94.17	94.93	94.39	94.98	94.36
	γ_{21}	94.87	94.24	94.91	94.28	94.96	94.36
$(30, 30, 60)$	γ_{11}	94.82	94.25	94.91	94.32	94.80	94.27
	γ_{21}	94.83	94.27	94.80	94.23	94.78	94.18
Sample Size: 600		**Coverage Probability**					
Coefficient of Variation (%)	**Parameter**	**Correlation Independence**		**Correlation 0.3**		**Correlation 0.7**	
		Model Based	**Robust**	**Model Based**	**Robust**	**Model Based**	**Robust**
$(5, 5, 20)$	γ_{11}	94.88	94.72	94.86	94.71	94.92	94.74
	γ_{21}	94.88	94.72	94.84	94.65	94.96	94.81
$(30, 30, 60)$	γ_{11}	95.01	94.83	95.05	94.86	95.01	94.82
	γ_{21}	94.96	94.83	95.04	94.84	95.07	94.94

Table 3.11: Table of Coverage Probabilities achieved under the Generalized Wedderburn Approach using both model-based and robust variance estimators with samples of size 60, 180 and 600 under three different correlations and a simulation of size 10^5

the generalized Wedderburn approach compares with Aitchison's approach in modeling this dataset.

No.	Sand	Silt	Clay	Depth	No.	Sand	Silt	Clay	Depth
1	0.78	0.20	0.03	10.40	21	0.10	0.54	0.37	47.10
2	0.72	0.25	0.03	11.70	22	0.17	0.48	0.35	48.40
3	0.51	0.36	0.13	12.80	23	0.10	0.55	0.34	49.40
4	0.52	0.41	0.07	13.00	24	0.05	0.55	0.41	49.50
5	0.70	0.26	0.04	15.70	25	0.03	0.45	0.52	59.20
6	0.66	0.32	0.01	16.30	26	0.11	0.53	0.36	60.10
7	0.43	0.55	0.02	18.00	27	0.07	0.47	0.46	61.70
8	0.53	0.37	0.10	18.70	28	0.07	0.50	0.43	62.40
9	0.15	0.54	0.30	20.70	29	0.04	0.45	0.51	69.30
10	0.32	0.41	0.27	22.10	30	0.07	0.52	0.41	73.60
11	0.66	0.28	0.06	22.40	31	0.05	0.49	0.46	74.40
12	0.70	0.29	0.01	24.40	32	0.04	0.48	0.47	78.50
13	0.17	0.54	0.29	25.80	33	0.07	0.52	0.41	82.90
14	0.11	0.70	0.20	32.50	34	0.07	0.47	0.46	87.70
15	0.38	0.43	0.19	33.60	35	0.07	0.46	0.47	88.10
16	0.11	0.53	0.36	36.80	36	0.06	0.49	0.45	90.40
17	0.18	0.51	0.31	37.80	37	0.06	0.54	0.40	90.60
18	0.05	0.47	0.48	36.90	38	0.02	0.48	0.49	97.70
19	0.16	0.50	0.34	42.20	39	0.02	0.48	0.50	103.70
20	0.32	0.45	0.23	47.00					

Table 3.12: The Arctic Lake Dataset

The Arctic Lake data is made up of 39 compositions of sediments samples recorded at different water depths (in metres). The compositional variables are sand, silt and clay and the corresponding compositions give the proportion of the three constituents by weight.

Prior to proceeding with showing results obtained with the two different modeling strategies, a ternary diagram of the compositional variables will be presented so as to obtain a better understanding of the data being analyzed.

Figure 3.1 gives the ternary diagram of the compositions in the Arctic Lake dataset in relation to the depth at which the samples have been taken. The ternary diagram is a unit-length equilateral triangle which provides a convenient way of presenting 3-part compositions in a plot. A composition whose proportions are all the same will give rise to a point in the ternary diagram with equal distances from the sides opposite the three vertices. Consider for example one of the compositions in the Arctic Lake dataset, $(0.704, 0.29, 0.006)$ for the components sand, silt and clay respectively. Such a point would be at a distance of 0.704 from the side opposite the vertex for sand, 0.29 from the side opposite the vertex for silt and 0.006 from the side opposite the vertex for clay. In Figure 3.1, this point is the one closest to the bottom edge joining sand and silt. Despite having some of the proportions close to zero, the Arctic lake dataset has no zeros.

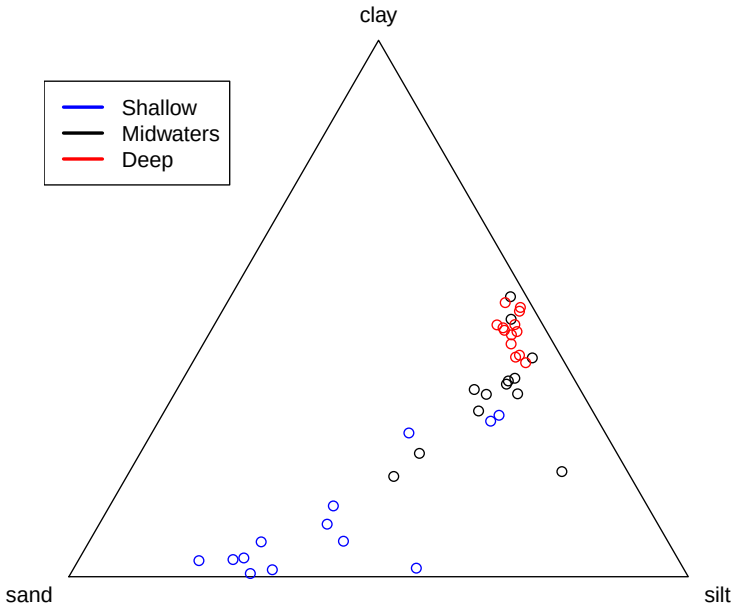


Figure 3.1: A Ternary Diagram showing Arctic Lake Compositional Data in Relation to Depth

The depths at which the samples were taken vary from 10.4m to 103.7m with an average of 48.04m. The relationship of the compositions with Depth in Figure 3.1 is expressed by using three different colours for the compositions: blue, black and red. The compositions marked blue are those which have been obtained at a level below the 33rd percentile of Depth. The compositions marked black are those which have been obtained at a level between the 33rd percentile and the 67th percentile of Depth. The remaining compositions, obtained at the deepest water levels, are marked in red. From Figure 3.1 it may be noticed that those samples that were obtained at the deepest water levels (marked in red) all contained a very small proportion of sand, giving a good indication of a relationship between Depth and the compositions.

Having obtained an initial feel for the data, we may now proceed to analyze the results obtained by using Aitchison's regression model and the generalized Wedderburn model. Results for Aitchison's regression model and the generalized Wedderburn model are obtained using the newly developed *cglm* package (see Chapter 5). Clay is taken as the reference component in the analysis and both models are fitted using log(Depth) as explanatory variable as per Aitchison (1986, p. 165).

The estimates obtained from using the two different approaches on the Arctic Lake dataset

are presented in Table 3.13. Recall from Section 3.4 that the two approaches estimate different mean models. Yet, from Table 3.13, it may be noticed that the estimates of the model coefficients obtained under the two methods are actually quite similar to each other. Also, the two sets of standard errors achieved using the generalized Wedderburn approach are not appreciably different from each other.

	Aitchison		**Generalized Wedderburn**		
Parameters	**Estimates**	**Standard Error**	**Estimates**	**Model-Based Standard Error**	**Robust Standard Error**
Intercept_1	9.697	1.004	8.665	0.764	0.737
Intercept_2	4.805	0.623	3.789	0.404	0.468
Log Depth_1	-2.743	0.269	-2.477	0.205	0.181
Log Depth_2	-1.096	0.167	-0.864	0.108	0.113

Table 3.13: Table of Estimates and their Standard Errors Obtained using Aitchison's approach and the Generalized Wedderburn approach on the Arctic Lake Dataset

For a further comparison of the performance of the two methods, the two measures of fit described in Section 3.6 have been applied to the Arctic Lake dataset. The resulting distance measures are presented in Table 3.14. The distance measure that is more in line with the generalized Wedderburn approach shows that the generalized Wedderburn approach provides a better fit (6.88) to the Arctic Lake dataset than Aitchison's approach (8.15). As expected, Aitchison's distance measure is slightly more in favour of the fit provided by Aitchison's regression model (8.94).

		Distance Measure	
		Aitchison	**Generalized Wedderburn**
Model	**Aitchison**	8.94	8.15
	Generalized Wedderburn	9.15	6.88

Table 3.14: Table of Distance Measures achieved using Aitchison's approach and the Generalized Wedderburn approach on the Arctic Lake Dataset

A ternary diagram with fitted values obtained for each method is shown in Figure 3.2. The fitted values obtained under the generalized Wedderburn approach are closer to most of the Arctic Lake compositions, with the fitted line achieved under Aitchison's approach seemingly being pulled towards the compositions with parts close to zero.

Using Cook's distance with a threshold of $4/n = 0.103$, three compositions have been identified as being influential under Aitchison's approach; compositions 6, 12 and to a lesser extent 18. The influential points are marked in red in the two plots in Figure 3.3, with compositions 6 and 12 being tagged in both plots. Compositions 6 and 12 correspond to the leftmost two points at the bottom of the ternary diagram (refer to Figure 3.2). These two compositions have a proportion of clay that is very close to zero.

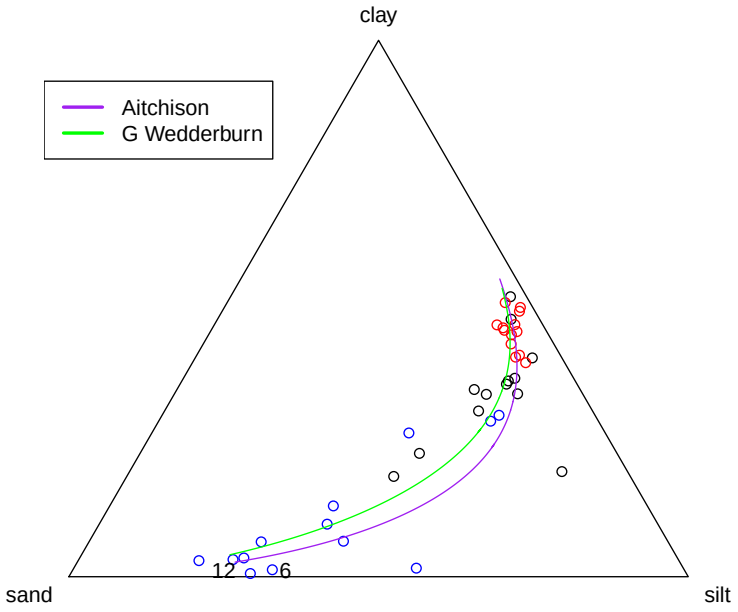


Figure 3.2: A Ternary Diagram showing the fitted lines achieved under Aitchison's approach and Generalized Wedderburn approach using the Arctic Lake Dataset

The generalized Wedderburn residuals and Aitchison residuals (see Section 3.6) are also computed for all three constituents. The residuals obtained for each method give three marginal views of a two-dimensional residual vector, so there is some redundancy. We are showing plots for all three constituents for each method for reasons of symmetry. Figures 3.4 and 3.5 display the generalized Wedderburn residuals and Aitchison residuals plotted against Log Depth. The resulting plots reveal no signs of mean-model misspecification and no sign of heteroscedasticity. Composition 12 however stands out once again from the rest by giving a slightly lower Aitchison residual value (-2.14) on the constituent Clay. This composition is marked in red in the respective figure.

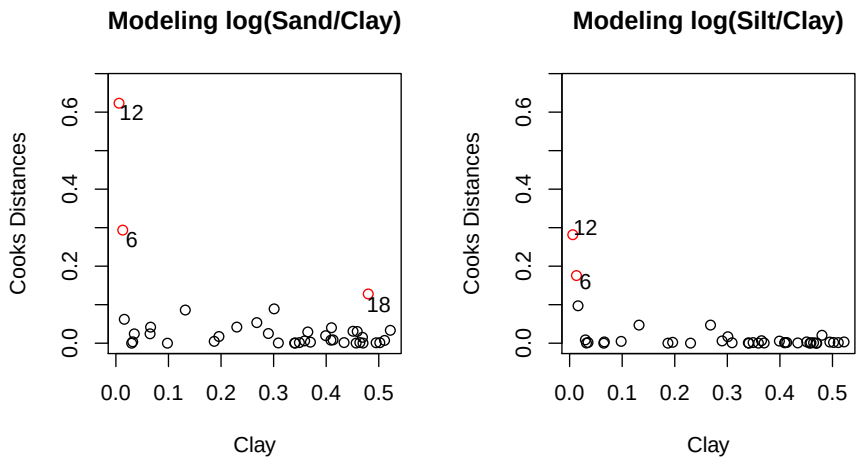


Figure 3.3: Cook's Distances Plots obtained using Aitchison's approach for the Arctic Lake Dataset

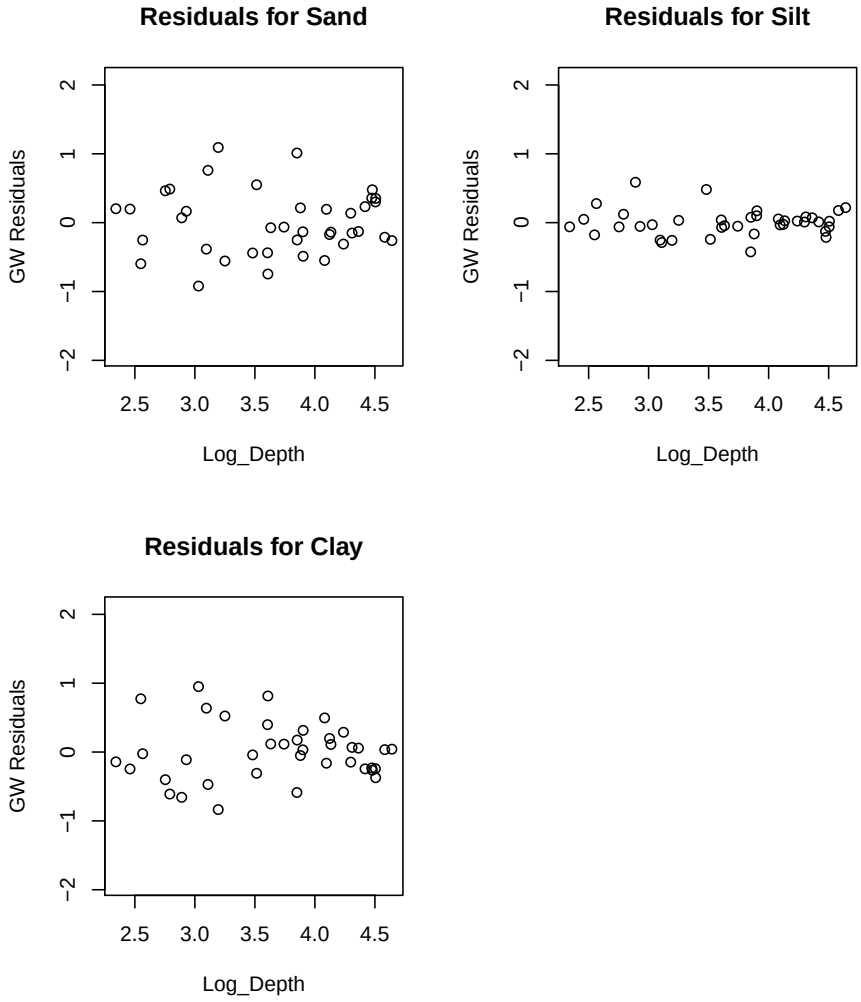


Figure 3.4: Plots of Generalized Wedderburn Residuals fitted against Log Depth for the Arctic Lake Dataset

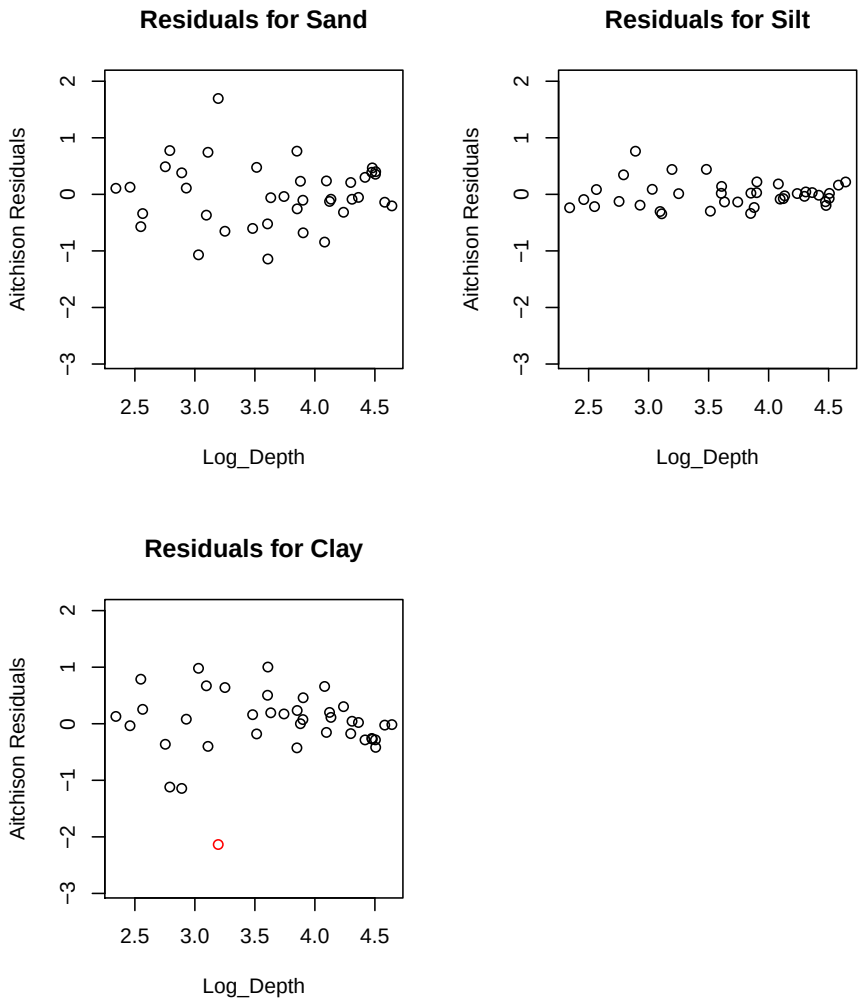


Figure 3.5: Plots of Aitchison Residuals fitted against Log Depth for the Arctic Lake Dataset

The Shapiro-Wilk test was performed on all six sets of residuals. The p-values obtained for the generalized Wedderburn residuals for Sand and Clay, 0.77 and 0.71 respectively, show that we do not have strong evidence against the two sets of residuals being normally distributed. The Aitchison residuals corresponding to Sand are also reasonably approximated by the normal distribution with a p-value of 0.46. On the other hand, there is enough evidence to reject normality for the remaining sets of residuals. The normal QQ plots obtained for the generalized Wedderburn residuals and Aitchison residuals (see Figures 3.6 and 3.7) in fact do show deviation from normality for the two sets of residuals pertaining to Silt and strong deviations from normality for the Aitchison residuals obtained for Clay. The three points at the bottom left hand corner of the QQ plot for Clay correspond to compositions 6, 7 and 12 with composition 12 being the composition closest to the bottom left hand corner. By looking at the Arctic Lake dataset (see Appendix 3.12), it may be noted that Compositions 6, 7 and 12 all have a relatively low value of Clay.

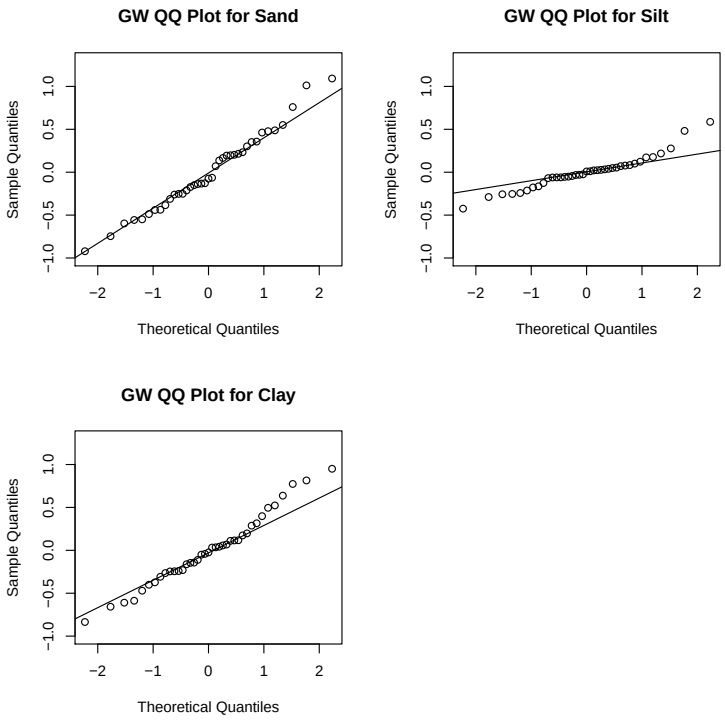


Figure 3.6: Normal QQ Plot of Generalized Wedderburn Residuals for the Arctic Lake Dataset

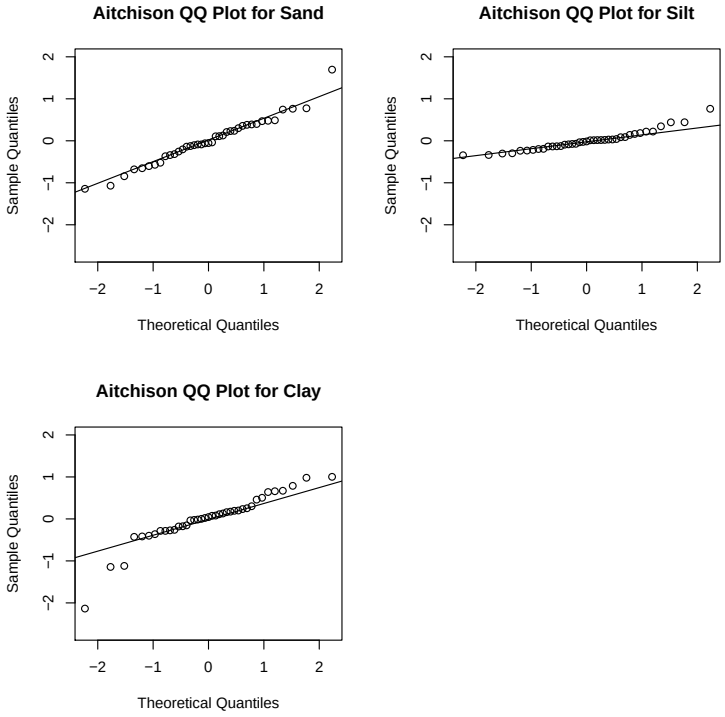


Figure 3.7: Normal QQ Plot of Aitchison Residuals for the Arctic Lake Dataset

3.10 Analyzing the Foraminiferal Dataset

The Foraminiferal dataset (see Table 3.15, Data 34 from Appendix D of Aitchison (1986)) is another dataset that is widely used in compositional data literature (e.g. Aitchison, 1986; Palarea-Albaladejo et al., 2007; Scealy and Welsh, 2011; Tsagris, 2015) particularly because this dataset contains zeros. As mentioned before, the generalized Wedderburn approach may be used even if there are zeros in the data. Aitchison's logratio approach may only be used if the zeros in the data are considered to be rounded zeros (see Section 1.3.1). If so, the zeros in the data may be replaced by imputed values prior to the logratio transformation.

In this section we will assume that the zeros in the Foraminiferal dataset are rounded zeros. The modified EM algorithm (Palarea-Albaladejo et al., 2007; Palarea-Albaladejo and Martín-Fernández, 2008) will be used to impute the zeros in the data by means of the function *impRZalr* in the R package *robCompositions*. Aitchison's regression model will then be fitted to the imputed dataset. The generalized Wedderburn approach will be used on the dataset with zeros as well as the imputed dataset. This example will serve as a further illustration of how Aitchison's approach compares with the generalized Wedderburn approach and will also serve as a kind of sensitivity analysis for the generalized Wedderburn approach.

The Foraminiferal data is made up of 30 compositions of foraminifer, a single-celled marine microorganism, with the compositions being recorded at different water depths (in metres) with depth varying from 1m to 30m. The compositional variables considered in this dataset are Neogloboquadrina atlantica (Na), Neogloboquadrina pachyderma (Np), Globorotalia obesa (Go) and Globigerinoides triloba (Gt). Five of the compositions contain a zero value for either Go or Gt. As per Scealy and Welsh (2011), it is possible that the proportions recorded for observation 24 on Go and Gt 'have been swapped by mistake' (see Table 3.15). We agree with such a remark. In this analysis we will thus proceed as suggested by Scealy and Welsh (2011) and swap the values for Go and Gt for observation 24.

Ternary diagrams for this dataset are presented in Figure 3.8. Since the number of parts in the foraminiferal compositions is four, a matrix of ternary diagrams for different sub-compositions is displayed. As for the Arctic lake dataset, the colour chosen for each composition in the ternary diagrams is in accordance with Depth. The compositions marked blue are those which have been obtained at a level below the 33rd percentile of Depth. The compositions marked black are those which have been obtained at a level between the 33rd percentile and the 67th percentile of Depth. The remaining compositions, obtained at the deepest water levels, are marked in red. From the ternary diagrams involving the constituent Gt, that is the ternary diagram at the top right and the two ternary diagrams at the bottom, it may be noticed that there seems to be a tendency for small values of Gt to be obtained at the deepest water levels, giving an indication of a relationship between the compositions and Depth. This relationship may also be corroborated by looking at

the last two columns of the data in Table 3.15.

	Proportions				
Sample Number	**Na**	**Np**	**Go**	**Gt**	**Depth (in metres)**
1	0.74	0.19	0.03	0.04	1
2	0.58	0.29	0.01	0.12	2
3	0.58	0.19	0.22	0.01	3
4	0.61	0.28	0.08	0.03	4
5	0.82	0.13	0.02	0.03	5
6	0.48	0.38	0.01	0.13	6
7	0.59	0.38	0.00	0.03	7
8	0.76	0.12	0.09	0.03	8
9	0.81	0.12	0.04	0.03	9
10	0.68	0.23	0.05	0.04	10
11	0.72	0.20	0.04	0.04	11
12	0.62	0.27	0.09	0.02	12
13	0.45	0.25	0.29	0.01	13
14	0.66	0.25	0.06	0.03	14
15	0.85	0.13	0.01	0.01	15
16	0.75	0.09	0.15	0.01	16
17	0.69	0.25	0.00	0.06	17
18	0.76	0.10	0.11	0.03	18
19	0.66	0.29	0.01	0.04	19
20	0.66	0.24	0.06	0.04	20
21	0.50	0.46	0.00	0.04	21
22	0.65	0.25	0.05	0.05	22
23	0.60	0.35	0.02	0.03	23
24	0.40	0.27	0.01	0.32	24
25	0.60	0.10	0.30	0.00	25
26	0.60	0.10	0.29	0.01	26
27	0.59	0.39	0.01	0.01	27
28	0.58	0.39	0.01	0.02	28
29	0.61	0.34	0.02	0.03	29
30	0.39	0.49	0.12	0.00	30

Table 3.15: The Foraminiferal Dataset with Proportions: Na: Neogloboquadrina Atlantica, Np: Neogloboquadrina Pachyderma, Go: Globorotalia Obesa, Gt: Globigerinoides Triloba

Prior to proceeding with showing results obtained with the two different modeling strategies, some detail on the modified EM algorithm is given. As for the standard EM algorithm, the modified EM algorithm is based on a complete dataset augmented with missing data. The alr transformed raw data without zeros makes up the complete part of the data. The zeros in the data are defined as missing data that needs to be imputed. The first component (Na) is chosen as reference for the Foraminiferal dataset. As for the standard EM algorithm, the complete data is assumed to follow a multivariate normal distribution with some mean vector and variance-covariance matrix, and the imputation procedure

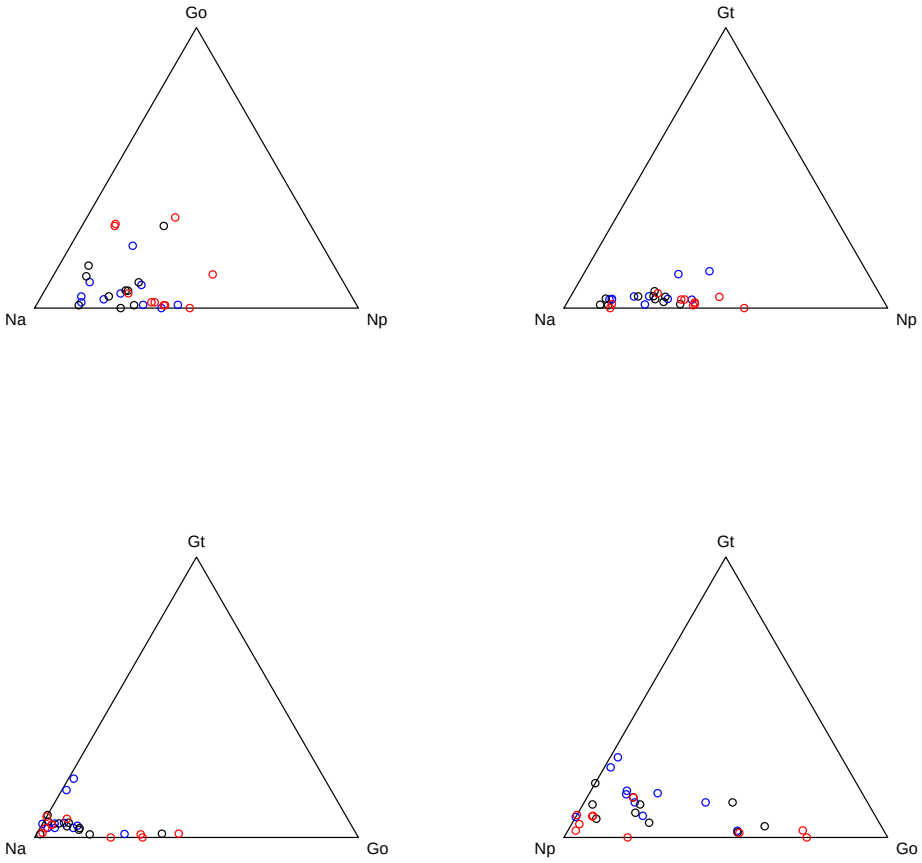


Figure 3.8: A Matrix of Ternary Diagrams for the Foraminiferal Subcompositions in Relation to Depth

involves the two steps of expectation and maximization. In the modified EM algorithm, however, the expectation step is modified so as to take into account of the alr-transformed detection limit of the compositions. Once a completed dataset is obtained, the data is back transformed to the simplex by means of the inverse alr transformation.

The reasons for choosing the modified EM algorithm for imputing the Foraminiferal dataset are various. As per Palarea-Albaladejo et al. (2007), the technique caters for the fact that the imputed values must be lower than the given detection limit, it is independent of the selected divisor in the alr transformation and 'the covariance structure of the parts without zeros is preserved' following imputation. It is also 'coherent with the basic operations and the vector space structure of the simplex', that is, it is subcomposition invariant, perturbation invariant and power transformation invariant. For more detail on these invariance properties refer to Palarea-Albaladejo et al. (2007, p. 633). The modified EM algorithm particularly outperforms other imputation techniques used for compositional data when the number of zeros is large. It may however also be used in

the presence of small number of zeros, as is the case with the Foraminiferal dataset.

We can now move onto fitting Aitchison's regression model to the imputed dataset and to use the generalized Wedderburn approach with both raw and imputed dataset. Following Scealy and Welsh (2011) and Tsagris (2015), modelling of this data is carried out by considering Depth as the explanatory variable. The estimates that resulted from fitting the three models are presented in Tables 3.16 and 3.17.

	Generalized Wedderburn Without Imputation		
Parameters	**Estimates**	**Model-Based Standard Error**	**Robust Standard Error**
Intercept_1	2.430	0.284	0.311
Intercept_2	1.170	0.258	0.252
Intercept_3	-0.151	0.635	0.643
Depth_1	0.039	0.016	0.016
Depth_2	0.061	0.015	0.016
Depth_3	0.074	0.036	0.036

Table 3.16: Table of Estimates and their Standard Errors Obtained using the Generalized Wedderburn approach on the Foraminiferal Dataset Without Imputation

	Aitchison Estimates		**Generalized Wedderburn**		
Parameters	**Estimates**	**Standard Error**	**Estimates**	**Model-Based Standard Error**	**Robust Standard Error**
Intercept_1	2.734	0.270	2.472	0.271	0.312
Intercept_2	1.420	0.254	1.212	0.239	0.244
Intercept_3	-0.333	0.702	-0.099	0.616	0.638
Depth_1	0.032	0.015	0.035	0.015	0.015
Depth_2	0.051	0.014	0.056	0.013	0.014
Depth_3	0.052	0.040	0.069	0.035	0.035

Table 3.17: Table of Estimates and their Standard Errors Obtained using Aitchison's approach and the Generalized Wedderburn approach on the Foraminiferal Dataset With Imputation

From the results in Tables 3.16 and 3.17 it may be noticed that there are only slight differences in the estimates of the model coefficients obtained using the generalized Wedderburn approach for the data with imputation and without. Minor differences may also be noticed in the resulting model-based and robust standard errors obtained under the generalized Wedderburn approach for the data with and without imputation. The fact that the results are so similar could be the result of having only five zeros in the dataset. Although the robust standard errors are more variable in repeated sampling, for this particular dataset, they are not appreciably different from the model-based ones. The estimates of the model coefficients and the corresponding standard errors obtained using Aitchison's approach are also similar to those obtained under the generalized Wedderburn method.

The two measures of fit described in Section 3.6 have also been applied to the Foraminiferal dataset. The resulting distance measures are presented in Table 3.18. The distance measure that is more in line with the generalized Wedderburn approach strongly favours the generalized Wedderburn approach. By looking at $y_{ij}/\hat{p}_{ij} - 1$ for all i and for all j in the Foraminiferal dataset, we find that the main contributors for such a relatively large distance measure are compositions 3, 13, 24, 25, 26 and to a lesser extent composition 16. By looking at the data in Table 3.15, it may be noticed that these compositions all have relatively large values of the component Go and small values in the component Gt, in comparison to the other compositions.

		Distance Measure	
		Aitchison	**Generalized Wedderburn**
Model	**Aitchison**	16.41	25.03
	Generalized Wedderburn Imputed	17.52	14.33
	Generalized Wedderburn Not Imputed	NA	14.67

Table 3.18: Table of Distance Measures achieved using Aitchison's approach and the Generalized Wedderburn approach on the Foraminiferal Dataset

From Table 3.18 it may also be noticed that there is barely any difference in the generalized Wedderburn distance measures obtained for the generalized Wedderburn approach using data with (14.33) or without imputation (14.67). This similarity follows from the fact that the analysis of the two datasets led to very similar results. As expected, Aitchison's distance measure favours the fit provided by Aitchison's regression model (16.41). Aitchison's distance obtained when the generalized Wedderburn approach is used with imputed data (17.52) is only slightly larger than the distance achieved when Aitchison's regression model is used (16.41). Since Aitchison's measure is based on logratios, it could not be computed without imputing the data.

The generalized Wedderburn residuals and Aitchison residuals (see Section 3.6) have also been computed for all the four parts for the imputed dataset. Figures 3.9 and 3.10 display the two sets of residuals plotted against Depth. The resulting plots reveal no signs of mean-model misspecification and no sign of heteroscedasticity. Compositions 3 and 13, however, give slightly higher generalized Wedderburn residuals (2.36 and 2.32 respectively) than the remaining compositions, on the component Go. These two residuals are marked in red in the respective plot in Figure 3.9.

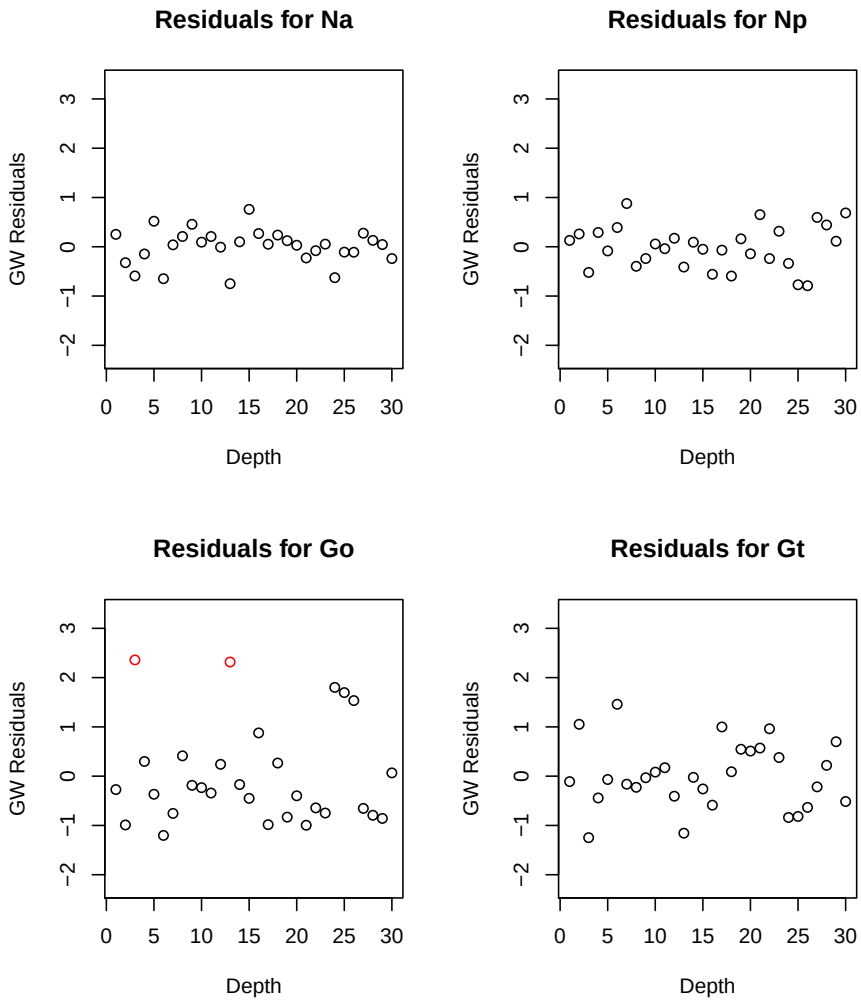


Figure 3.9: Plots of Generalized Wedderburn Residuals fitted against Depth for the Foraminiferal Dataset

The Shapiro-Wilk test was performed on all eight sets of residuals obtained for the imputed dataset. The resulting p-values show that all sets of Aitchison residuals and all sets of generalized Wedderburn residuals, except for generalized Wedderburn residuals for Go (p-value 0.0009) are reasonably approximated by the normal distribution. The normal QQ plots obtained for the generalized Wedderburn residuals for Go do in fact show strong deviations from normality, with the main contributors being compositions 3, 6, 13, 16, 21, 24, 25 and 26. On looking at the Foraminiferal dataset (see Table 3.15), it may be noticed that in comparison with the other values of Go, these compositions have either a very low or a very high value of Go.

In actual fact, neither the generalized Wedderburn model nor Aitchison's regression model performs very well in explaining the variability in the Foraminiferal compositions. Figures 3.13 and 3.14 display the fitted lines obtained for the generalized Wedderburn method and Aitchison's method when focusing on two subcompositions. Figure 3.13 shows the ternary

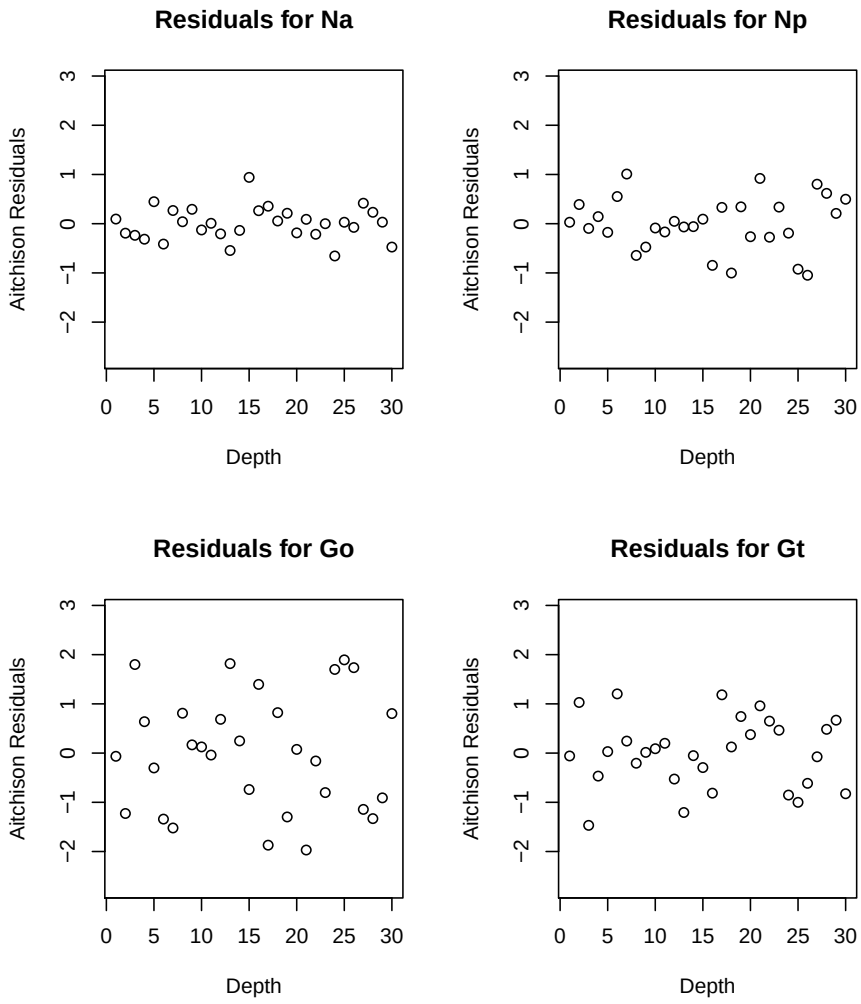


Figure 3.10: Plots of Aitchison Residuals fitted against Depth for the Foraminiferal Dataset

diagram with fitted lines for Np, Go and Gt and Figure 3.14 shows the ternary diagram with fitted lines for Na, Np and Go. From these figures, it may be noticed that the fitted values obtained through the generalized Wedderburn model and Aitchison's regression model differ quite substantially and the fit obtained by each method is very poor. The percentage variability explained in the compositions by including the variable Depth in the model using either the generalized Wedderburn model or Aitchison's regression model are in fact 18% and 6% respectively, which are both very low. These percentages have been obtained by computing $1-205.32/249.87 = 0.18$ and $1-269.27/285.07 = 0.06$, where 205.32 and 249.87 are the squared generalized Wedderburn distance measures obtained by using the generalized Wedderburn approach with Depth and without Depth in the model, and 269.27 and 285.07 are the squared Aitchison distance measures obtained by using Aitchison's regression model with Depth and without Depth in the model.

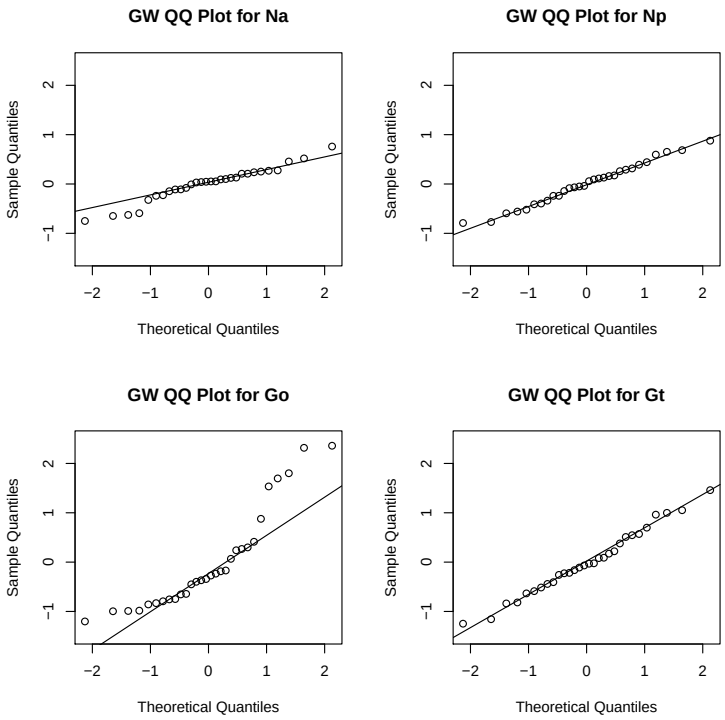


Figure 3.11: Normal QQ Plot of Generalized Wedderburn Residuals for the Foraminiferal Dataset

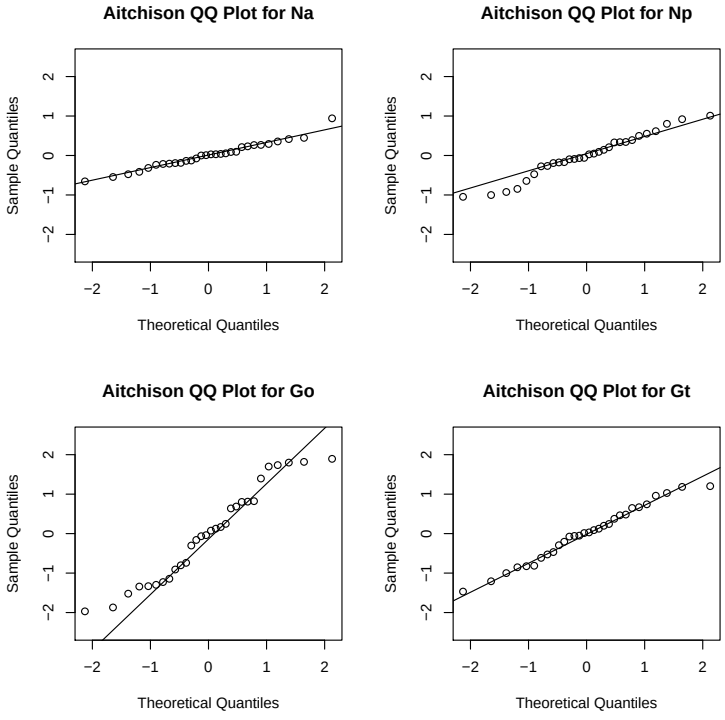


Figure 3.12: Normal QQ Plot of Aitchison Residuals for the Foraminiferal Dataset

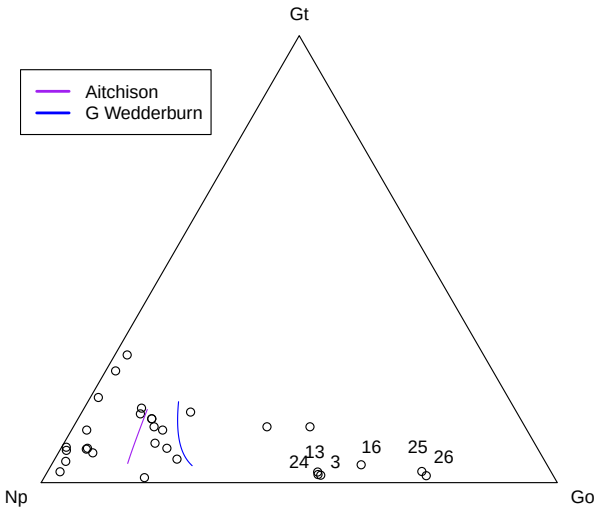


Figure 3.13: A Ternary Diagram showing the Fitted Lines achieved for the Subcompositions of Np, Go and Gt, using the Generalized Wedderburn Model and Aitchison's Model for the Foraminiferal Dataset

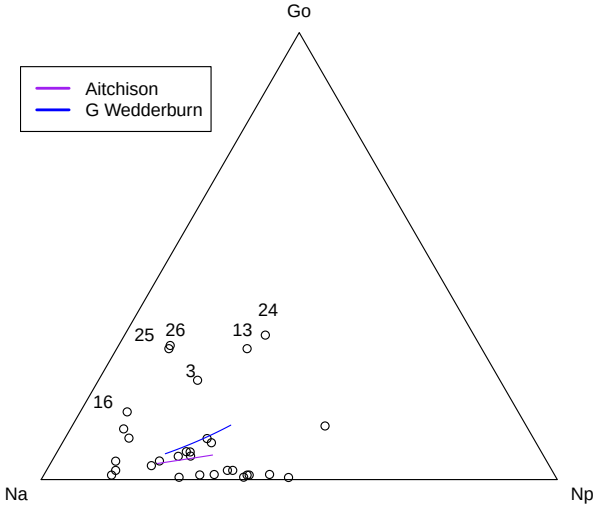


Figure 3.14: A Ternary Diagram showing the Fitted Lines achieved for the Subcompositions of Na, Np and Go, using the Generalized Wedderburn Model and Aitchison's Model for the Foraminiferal Dataset

Chapter 4

Further Empirical Study of the Generalized Wedderburn Method

As has been mentioned in Section 1.2.1, the Dirichlet distribution is one of the familiar classes of distributions that has been used to model continuous compositional data. Problems faced when using a Dirichlet distribution to model the influence of explanatory variables on compositional response variables have also been mentioned in Section 1.2.1. The Dirichlet regression model may however be specified using the same logit model that is estimated by the generalized Wedderburn approach. Also, the parameters in a Dirichlet regression model are estimated using maximum likelihood estimation. These two properties make the Dirichlet model attractive for use in a simulation study which compares the efficiency of the GEE estimator, used under the generalized Wedderburn approach, with the maximum likelihood estimator used in the Dirichlet regression model.

Theoretical background on the Dirichlet regression model will be given in Section 4.1. In Section 4.2, a Dirichlet regression model will be fitted to the Arctic Lake dataset (Data 5 from Appendix D of Aitchison (1986)). In Section 4.3, the setup for a simulation study based on the estimates obtained in Section 4.2 for the Arctic Lake dataset is presented. The results obtained from the simulation study are then discussed in Section 4.4.

4.1 The Dirichlet Regression Model

Detail on the probability density function of a Dirichlet distributed random vector $\mathbf{Y}$ together with the mean and the variance-covariance structure underlying the family of Dirichlet distributions has been provided in Section 1.2.1. The detail provided in Section 1.2.1 pertain to what is known as the 'common parametrization' of the Dirichlet distribution. So as to be able to compare the estimates obtained from a Dirichlet Regression model with those obtained from the generalized Wedderburn method, focus will be directed towards the 'alternative parametrization' of the Dirichlet distribution (Maier,

2014).

In the alternative parametrization, a new set of parameters p_j, $(j = 1, \ldots, J)$, are defined such that

$$E\left(Y_j\right) = p_j \tag{4.1}$$

and the parameter α_+ (defined as in Section 1.2.1) models the precision in the model. A high precision 'centres the density around the expected' mean vector whilst a low precision pushes the distribution of the points 'towards the sides and corners of the simplex' $\mathbb{S}^{J-1}$ (Maier, 2014).

The conversion of p_j back into the common parametrization in terms of α_j, is carried out using

$$\alpha_j = \alpha_+ p_j. \tag{4.2}$$

The probability density function of a Dirichlet distributed random vector $\mathbf{Y}$ under the alternative parametrization is thus given by

$$f\left(y_1, \ldots, y_{J-1} | \mathbf{p}, \alpha_+\right) = \frac{\Gamma\left(\sum_{j=1}^{J} \alpha_+ p_j\right)}{\prod_{j=1}^{J} \Gamma\left(\alpha_+ p_j\right)} \prod_{j=1}^{J} y_j^{\alpha_+ p_j - 1}, \tag{4.3}$$

where $\mathbf{p} = \left(p_1, \ldots, p_J\right)^{'}$.

The variance of Y_j under the alternative parametrization becomes:

$$\mathrm{Var}\left(Y_j\right) = \frac{p_j\left(1 - p_j\right)}{\alpha_+ + 1}, \tag{4.4}$$

and for $j \neq j^{'}$

$$\mathrm{Cov}\left(Y_j, Y_{j^{'}}\right) = -\frac{p_j p_{j^{'}}}{\alpha_+ + 1}. \tag{4.5}$$

Then, fitting of a Dirichlet regression model under the alternative parametrization is carried out by specifying a set of equations for the parameters p_j and an equation for the precision parameter α_+. Taking the last component as reference component, in a Dirichlet regression model based on the alternative parametrization, the means are modeled using

$$p_{ij} = \begin{cases} \frac{\exp\left(\mathbf{x}_i^{'} \boldsymbol{\gamma}_j\right)}{\sum_{j^{'}=1}^{J} \exp\left(\mathbf{x}_i^{'} \boldsymbol{\gamma}_{j^{'}}\right)} & \text{if } j = 1, \ldots, J-1 \\ \frac{1}{\sum_{j^{'}=1}^{J} \exp\left(\mathbf{x}_i^{'} \boldsymbol{\gamma}_{j^{'}}\right)} & \text{if } j = J. \end{cases} \tag{4.6}$$

where $\boldsymbol{\gamma}_j$ are the model coefficients that need to be estimated and $\mathbf{x}_i$ denotes the vector of observations obtained by the i^{th} case on explanatory variables $X_1, \ldots, X_p$. The logit link is chosen so as to ensure that the fitted means $\widehat{p}_{ij}$ are sum-constrained to 1 for each i. For the purpose of the analysis that will be carried out in the sections which follow, α_+

is modeled by an intercept only model leading to the same estimate of α_+ for all i.

So the multinomial logit strategy used to fit a Dirichlet regression model is the same as that used under the generalized Wedderburn approach (see equation (2.56)). Three main differences in the estimates achieved from the two different models may however be pointed out:

- Estimates in the Dirichlet regression model are obtained using maximum likelihood estimation whilst estimates in the generalized Wedderburn approach are obtained using generalized estimating equations. The desirable properties of maximum likelihood estimators are well known and understood. Through fitting a Dirichlet regression model and using the generalized Wedderburn approach on data simulated from a Dirichlet regression model it would be possible to compare how the efficiency of the GEE estimator fares in relation to the maximum likelihood estimator used in a Dirichlet regression model. Details on a simulation study which focuses on such an aspect are given in Sections 4.3 and 4.4.

- The Dirichlet regression model and the generalized Wedderburn model are based on two different variance-covariance structures. On using the generalized Wedderburn approach with data simulated through a Dirichlet model, it is to be expected that the model-based estimator $\widehat{\text{Var}\,(\widehat{\boldsymbol{\gamma}})}_M$ fares badly in relation to Liang and Zeger (1986) robust counterpart $\widehat{\text{Var}\,(\widehat{\boldsymbol{\gamma}})}_{LZ}$ (described in Section 2.8.3). Whilst the estimates obtained through GEE under the generalized Wedderburn approach are invariant to the choice of the 'working' variance-covariance structure (see Section 2.4.4), the model-based estimator $\widehat{\text{Var}\,(\widehat{\boldsymbol{\gamma}})}_M$ (2.67) relies on the estimation of $\phi\mathbb{V}_{\mathbf{p}_i,\boldsymbol{\Omega},\mathbb{W}}$ where $\phi\mathbb{V}_{\mathbf{p}_i,\boldsymbol{\Omega},\mathbb{W}}$ is the first order Taylor series approximation of the true variance-covariance matrix $\text{Var}\,(\mathbf{Y}_i)$ (see Section 2.7).

- Since the family of Dirichlet distributions is not a linear exponential family of distributions (Gourieroux et al., 1984), estimates achieved through a Dirichlet regression model are in general consistent only if there is no distributional misspecification. The GEE estimator is consistent provided that the marginal mean model specification is correct.

4.2 Fitting a Dirichlet Regression Model to the Arctic Lake Dataset

As mentioned in Section 3.9, the Arctic Lake dataset is a widely used dataset in compositional data literature making it an appealing dataset to use for comparison of various models that are typically used with compositional data. We have already analyzed the Arctic Lake dataset using Aitchison's approach and the generalized Wedderburn approach in Section 3.9, with log(Depth) as explanatory variable. A Dirichlet regression model under the alternative parametrization, also using log(Depth) as the explanatory variable and

$J = 3$ as reference component, will now also be fitted to the Arctic Lake dataset. The fitting of a Dirichlet model to the Arctic Lake dataset serves two purposes:

- to analyze how the fit of the Dirichlet model compares with the fit provided by Aitchison's approach and the generalized Wedderburn approach
- to use the resulting estimates of the model coefficients to perform a simulation study (details of which are presented in the subsequent section) with data generated from a Dirichlet model.

The Dirichlet model is fitted by means of the R package *DirichletReg* (Version 0.6-2). The estimates of the model coefficients and their standard errors are given in Table 4.1.

Parameters	**Estimates**	**Standard Error**
Intercept_1	8.39	0.69
Intercept_2	3.89	0.58
Log Depth_1	-2.38	0.19
Log Depth_2	-0.88	0.15

Table 4.1: Table of Estimates and their Standard Errors Obtained from fitting a Dirichlet Regression Model to the Arctic Lake Dataset

The estimates in Table 4.1 are similar to those achieved using the generalized Wedderburn approach (see Table 3.13). On the other hand, there is quite a difference between the estimates in Table 4.1 and those achieved from fitting Aitchison's method (see Table 3.13). This result was to be expected since the model fitted in Aitchison's approach is a model for different means from those fitted in the Dirichlet regression and the generalized Wedderburn method. On inspecting the ternary diagram in Figure 4.1 it may also be noted that the fitted lines resulting from the Dirichlet regression model and from the generalized Wedderburn model are closer to each other and seem to fit the majority of the compositions better than Aitchison's method does. The distance measures based on Aitchison's approach and the generalized Wedderburn approach have also been computed for the fitted Dirichlet regression model. The resulting two values are 9.21 and 6.89 respectively, which are very close to the values achieved under the generalized Wedderburn approach (see Table 3.14).

We shall now see how the maximum likelihood estimates, obtained from fitting a Dirichlet regression model to the Arctic Lake dataset, are used to generate data for a simulation study which compares the performance of the GEE estimator with the ML estimator after fitting a Dirichlet regression model.

4.3 The Simulation Setup

Due to the form of the probability density function of the Dirichlet distribution (1.2) (or equivalently (4.3)), the generation of a dataset from this distribution requires the

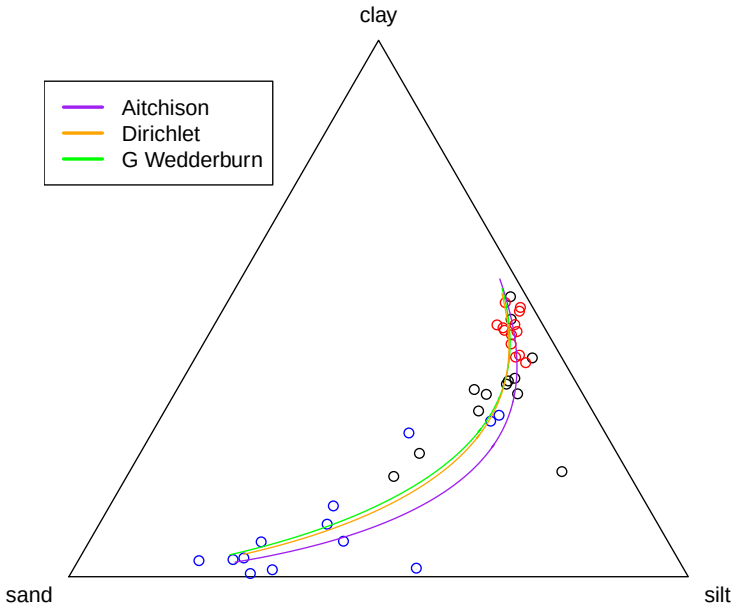


Figure 4.1: A Ternary Diagram showing the fitted Lines achieved under Dirichlet Regression, Aitchison's Method and the Generalized Wedderburn Method using the Arctic Lake Dataset

knowledge of the values of the parameters $\alpha_1, \ldots, \alpha_J$. The *DirichletReg* package facilitates the generation of data from a Dirichlet model by making the matrix of α estimates available as part of the output obtained after a Dirichlet regression model is fitted to a dataset. The generation of the 3-part compositional response variables used to perform the simulation study in this section, makes use of the α matrix that resulted after fitting the Dirichlet regression model to the Arctic Lake dataset and the specification of a sample of size 39. Recall that 39 is the sample size in the Arctic Lake dataset.

The Dirichlet model has been used to generate 100,000 sets (samples) of compositional response variables. The model (4.6) has been fit on every generated sample with log(Depth) (from the Arctic Lake dataset) as explanatory variable. Maximum likelihood estimates of the model coefficients have been obtained by means of the *DirichletReg* package. GEE estimates have been obtained by means of the package *cglm*.

Summarization of the Simulation Results

The estimates that are obtained at the end of the simulation are:

- the average of the resulting γ estimates achieved under both estimation techniques
- the biases achieved under the two techniques together with their standard error
- the variance of the γ estimates achieved under the two techniques together with the corresponding standard error
- the average of the estimated Var $(\widehat{\boldsymbol{\gamma}})$ using both model-based and robust variance estimators, under the generalized Wedderburn approach, together with their standard error

Since no special zero structure has been imposed in the model, the intercept parameter does not bear much interest. Focus will thus be directed towards the non-intercept parameters. So all the results obtained from the simulation study will focus on the coefficients γ_{11} and γ_{21}, the coefficients corresponding to the explanatory variable log(Depth).

4.4 Results obtained from the Simulation Study

The first aspect to be studied in this simulation study is the estimated bias that was obtained under the two estimation techniques. Results for the estimated bias together with the corresponding estimated standard errors are shown in Table 4.2.

	Dirichlet Estimates			**Generalized Wedderburn**		
Parameter $\times 10^2$	**Mean** $\times 10^2$	**Bias** $\times 10^2$	**Standard Error** $\times 10^2$	**Mean** $\times 10^2$	**Bias** $\times 10^2$	**Standard Error** $\times 10^2$
$\gamma_{11} = -238$	-239.376	-1.300	0.059	-241.897	-3.821	0.074
$\gamma_{21} = -88$	-88.433	-0.399	0.048	-89.844	-1.810	0.060

Table 4.2: Table of Estimates, Estimated Bias and Estimated Standard Error of the Bias achieved using MLE and GEE on Dirichlet simulated data based on estimates from the Arctic Lake dataset with a sample of size 39 and a simulation of size 10^5

From Table 4.2 it may be noticed that both estimation techniques led to some bias in the resulting estimates. On considering that the desirable property of unbiasedness of maximum likelihood estimators holds as $n \to \infty$, the bias obtained in this simulation study, based on a sample of size 39, was to be expected. The property of asymptotic unbiasedness of GEE estimators is well known (Liang and Zeger, 1986). Limited literature is however available on performance of GEE estimators when the sample size used is small. Paul et al. (2013) developed a bias-corrected GEE estimator based on the bias corrective measure used to correct the order $\frac{1}{n}$ bias for maximum likelihood estimators (see Cox and Snell, 1968). Paul et al. (2013) use the bias correction in conjunction with longitudinal binary response data and conclude that for small sample sizes, the bias-corrected GEE estimator 'shows superior performance in terms of bias and efficiency' in comparison to the standard GEE estimator. Paul and Zhang (2014) developed another bias-adjusted GEE estimator based

on a bias preventive measure used to correct the bias for maximum likelihood estimators (see Firth, 1993a). Using simulations, Paul and Zhang (2014) show that particularly for samples of size ≤ 50, the two bias-adjusted estimators show 'improvement in bias, mean square error, standard error and length of confidence intervals of the estimates'. The bias obtained by the GEE estimator in our simulation study when using a sample of size 39, seems to suggest that the bias-adjusted estimators used in Paul et al. (2013) and Paul and Zhang (2014) might be worth investigating when using the generalized Wedderburn approach with small sample sizes.

Also, the absolute values of the bias obtained by estimating the parameters using generalized estimating equations are larger than those obtained by using maximum likelihood estimation and from Table 4.3 it may be noticed that the estimated variances achieved under MLE are appreciably smaller than those achieved using GEE. The relative efficiencies of the two parameters are in fact 0.63 and 0.66 respectively, showing that the GEE estimator lost around 1/3 of the efficiency when compared to the maximum likelihood estimator in this simulation study. Since GEE estimators are obtained as a result of the specification of the marginal distribution of the compositional variables rather than through the specification of the joint likelihood function, it is however to be expected that the ML estimator fares better than the GEE estimator.

	Dirichlet Regression Estimates		**Generalized Wedderburn**	
Parameter $\times 10^2$	**Variance** $\times 10^2$	**Standard Error** $\times 10^2$	**Variance** $\times 10^2$	**Standard Error** $\times 10^2$
$\gamma_{11} = -238$	3.487	0.016	5.555	0.025
$\gamma_{21} = -88$	2.333	0.010	3.555	0.016

Table 4.3: Table of Variance Estimates together with their Estimated Standard Errors achieved using MLE and GEE on Dirichlet simulated data based on Estimates from the Arctic Lake dataset with a sample of size 39 and a simulation of size 10^5

Seeing that the Dirichlet model is based on a different variance-covariance structure from the one considered in the generalized Wedderburn approach, it is not surprising to see that the Liang and Zeger (1986) robust estimator fares better at estimating the variance of the GEE estimator than the model-based estimator in this case (refer to Table 4.4). From the results in Table 4.4 it may also be noticed that both the model-based and Liang and Zeger (1986) robust estimator exhibit some downward bias in estimating the actual variance of the estimators. Issues related to the use of Liang and Zeger (1986) robust estimator with small sample sizes have been mentioned in Section 2.8.3. The downward bias, in particular, achieved by the robust estimator in analyzing small sample sizes is well documented (e.g. Emrich and Piedmonte, 1992; Drum and McCullagh, 1993; Mancl and DeRouen, 2001; Pan, 2001b; Gosho et al., 2014).

Generalized Wedderburn Variance Estimates						
			Model-Based		Robust	
Parameter	**Variance** $\times 10^2$	**Standard Error** $\times 10^2$	**Mean of** $\widehat{\text{Var}(\widehat{\gamma})}_M$ $\times 10^2$	**Standard Error** $\times 10^2$	**Mean of** $\widehat{\text{Var}(\widehat{\gamma})}_{LZ}$ $\times 10^2$	**Standard Error** $\times 10^2$
γ_{11}	5.555	0.025	4.0523	0.0033	4.5952	0.0057
γ_{21}	3.555	0.016	2.2374	0.0020	2.9647	0.0047

Table 4.4: Table of Variance Estimates together with their Estimated Standard Errors achieved under the Generalized Wedderburn Approach using Dirichlet simulated data based on Estimates from the Arctic Lake dataset with a sample of size 39 and a simulation of size 10^5

Chapter 5

The *cglm* Software Package

5.1 Introduction

This chapter briefly describes an early, development version of a public package for the *R* statistical computing environment (RCore Team, 2016). The *cglm* package is the joint work of David Firth and Fiona Sammut. The package can be added to an *R* installation by

```
> devtools::install_bitbucket("davidfirth/cglm")
```

and made available in the standard way to the current *R* session by

```
> library(cglm)
```

At the time of writing this description, in September 2016, the package has fully functional but rudimentary facilities for specifying and fitting the Generalized Wedderburn and Aitchison multivariate regression models, as well as basic tools for model summary and model criticism. A more complete version of the package, which generalizes (to include compositional-response regressions) most of the capabilities of *R*'s standard `glm` function and associated methods, is planned for publication via the *Comprehensive R Archive Network* by early 2017.

5.2 Model Specification and Fitting

The core of the *cglm* package is the `cglm` function itself, whose operation mimics that of `glm` from *R*'s standard *stats* package. The full documentation for the `cglm` function can be found via `help(cglm)`. Here we illustrate the use of `cglm` to fit models to the 39 Arctic Lake sediment compositions given in Aitchison (1986). The data are provided in the *cglm* package as a 4-column matrix whose first three columns give the compositions; the fourth column is the `depth` variable.

```
> data(ArcticLake)
> head(ArcticLake)
```

```
  sand silt clay depth
1 77.5 19.5  3.0  10.4
2 71.9 24.9  3.2  11.7
3 50.7 36.1 13.2  12.8
4 52.2 40.9  6.6  13.0
5 70.0 26.5  3.5  15.7
6 66.5 32.2  1.3  16.3
```

The compositions all sum approximately to 100:

```
> sediments <- ArcticLake[, c("sand", "silt", "clay")]
> rowSums(sediments)

    1     2     3     4     5     6     7     8     9    10
100.0 100.0 100.0  99.7 100.0 100.0 100.0 100.0 100.0 100.0
   11    12    13    14    15    16    17    18    19    20
100.0 100.0 100.0 100.0 100.0 100.0 100.0 100.0 100.0 100.0
   21    22    23    24    25    26    27    28    29    30
100.0 100.0 100.0 100.5 100.0 100.0 100.0 100.0 100.0  99.9
   31    32    33    34    35    36    37    38    39
100.0 100.0 100.0  99.9  99.9 100.0 100.0 100.0 100.0
```

5.2.1 Generalized Wedderburn Model

The default action of the `cglm` function is to specify and fit a Generalized Wedderburn logit model. For the dependence of sediment composition upon log(depth), as summarized in Table 3.13:

```
> logdepth <- log(ArcticLake[, "depth"])
> gw_model <- cglm(sediments ~ logdepth, ref = 3)
```

The model-formula specification is as for `glm`. The additional `ref` argument specifies which component will be (arbitrarily) taken to be the reference component in the model; here the third component (`clay`) has been chosen. The default choice is `ref = 1`.

```
> gw_model

Call:
cglm(formula = sediments ~ logdepth, ref = 3)

Coefficients:
             sand     silt     clay
(Intercept)  8.6649   3.7890   0.0000
logdepth    -2.4767  -0.8642   0.0000
```

5.2.2 Aitchison-type Model

The corresponding multivariate linear regression of logratios (see also Table 3.13), again relative to the component specified via the `ref` argument, is achieved by using the additional argument `method = "logy.fit"`:

```
> Ait_model <- cglm(sediments ~ logdepth, ref = 3,
+                   method = "logy.fit")
> Ait_model

Call:
cglm(formula = sediments ~ logdepth, method = "logy.fit", ref = 3)


Coefficients:
             sand    silt    clay
(Intercept)   9.697   4.805   0.000
logdepth     -2.743  -1.096   0.000
```

The currently available options for the `method` argument are `"logy.fit"` as here for the multivariate linear regression with logratios of the data, and the default `"gw.fit"` which implements the Generalized Wedderburn model.

The fitted totals agree with those found in the data. For example

```
> ## Compare the fitted totals below with those found above for
> ## the sediments data
> round(rowSums(fitted(Ait_model)) - rowSums(sediments), 2)

 1  2  3  4  5  6  7  8  9 10 11 12 13 14 15 16 17 18 19 20
 0  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0
21 22 23 24 25 26 27 28 29 30 31 32 33 34 35 36 37 38 39
 0  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0
```

5.3 Standard Errors and the summary Method

The `vcov.cglm` method estimates the variance-covariance matrix of the regression parameter estimates. This is the usual model-based variance-covariance matrix. For the generalized Wedderburn model the robust, 'sandwich'-type estimator can be specified instead via the argument `type = "robust"`.

```
> vcov(gw_model)

                 (Intercept)_sand logdepth_sand
(Intercept)_sand       0.58425607   -0.15398423
logdepth_sand         -0.15398423    0.04194471
(Intercept)_silt       0.24152910   -0.06365646
```

```
logdepth_silt         -0.06365646    0.01733977
(Intercept)_clay       0.00000000    0.00000000
logdepth_clay          0.00000000    0.00000000
                 (Intercept)_silt logdepth_silt
(Intercept)_sand       0.24152910   -0.06365646
logdepth_sand         -0.06365646    0.01733977
(Intercept)_silt       0.16336482   -0.04305579
logdepth_silt         -0.04305579    0.01172823
(Intercept)_clay       0.00000000    0.00000000
logdepth_clay          0.00000000    0.00000000
                 (Intercept)_clay logdepth_clay
(Intercept)_sand                0             0
logdepth_sand                   0             0
(Intercept)_silt                0             0
logdepth_silt                   0             0
(Intercept)_clay                0             0
logdepth_clay                   0             0
```

The resulting standard errors, computed as square roots of diagonal entries in the variance-covariance matrix, are conveniently displayed through the `summary.cglm` method. For example, here with the 'robust' standard errors as shown in Table 3.13:

```
> summary(gw_model, vcov_type = "robust")

Call :
cglm(formula = sediments ~ logdepth, ref = 3)

Residuals :
           sand      silt     clay
Min    -0.92080 -0.424500 -0.83530
1Q     -0.28590 -0.065290 -0.24490
Median -0.07401  0.009065 -0.02435
3Q      0.26700  0.073970  0.18580
Max     1.09200  0.586900  0.95060

Coefficients :
   sand / clay :
             Estimate St. err
(Intercept)     8.665  0.7365
logdepth       -2.477  0.1809

   silt / clay :
             Estimate St. err
```

```
(Intercept)   3.7890  0.4684
logdepth     -0.8642  0.1128
```

5.4 Model Residuals

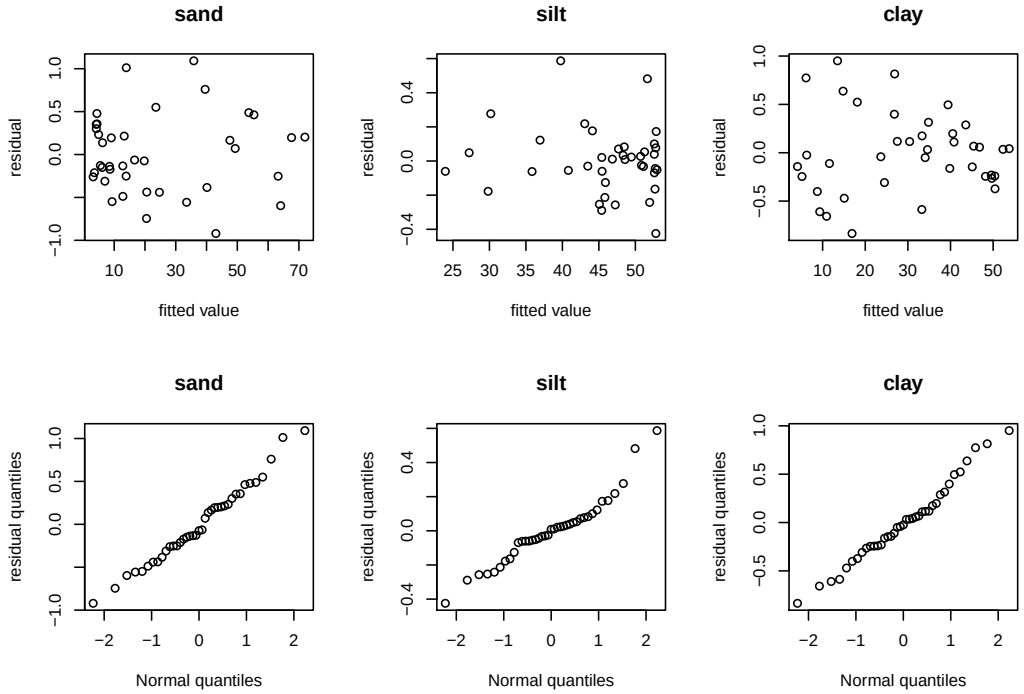
The result of applying `residuals()` to a `cglm` model object is a matrix with the same dimensions as the response matrix. The residual row sums are zero, and projections onto columns of the model matrix are all null:

```
> res <- residuals(gw_model)
> head(res)

        sand        silt        clay
1  0.2029363 -0.06036295 -0.14257334
2  0.1972738  0.04847321 -0.24574698
3 -0.5965236 -0.17730003  0.77382368
4 -0.2526709  0.27702057 -0.02434964
5  0.4626519 -0.06152271 -0.40112919
6  0.4878510  0.12220909 -0.61006010

> summary(rowSums(res))

      Min.    1st Qu.     Median       Mean    3rd Qu.
-1.665e-16  0.000e+00  0.000e+00 -2.135e-18  0.000e+00
      Max.
 8.327e-17

> crossprod(res, gw_model$x)

       (Intercept)      logdepth
sand  3.204064e-10  1.015552e-09
silt  1.873365e-10  3.557543e-10
clay -5.077430e-10 -1.371307e-09
```

These can be equivalently thought of as *working residuals* or *Pearson* (standardized) *residuals*. They are suitable for plotting for diagnostic purposes against such quantities as the fitted values, or candidate predictor variables not included in the model, or normal quantiles. For example:

```
> par(mfrow = c(2,3))
> fit <- fitted(gw_model)
> for (component in colnames(res)) {
+     plot(fit[, component], res[, component], xlab = "fitted value",
+          ylab = "residual", main = component)
+ }
```

```
> for (component in colnames(res)) {
+     qqnorm(res[, component], xlab = "Normal quantiles",
+            ylab = "residual quantiles", main = component)
+ }
```



The sum-to-zero property of the residuals needs to be kept in mind when looking at such plots: in this application, for example, there are just two residual degrees of freedom per observation, not three.

5.5 Still to be Added

In due course, the finished *cglm* package will include the full set of methods that correspond to standard methods available for `glm` objects. The current version of the package has error-generating place-holders for the most important of these:

```
plot.cglm <- function (x, ...) .NotYetImplemented()
predict.cglm <- function (object, ...) .NotYetImplemented()
add1.cglm <- function (object, scope, ...) .NotYetImplemented()
anova.cglm <- function (object, ...) .NotYetImplemented()
confint.cglm <- function (object, parm, level = 0.95, ...) {
    .NotYetImplemented() }
cooks.distance.cglm <- function (model, ...) .NotYetImplemented()
drop1.cglm <- function (object, scope, ...) .NotYetImplemented()
effects.cglm <- function (object, ...) .NotYetImplemented()
influence.cglm <- function (model, ...) .NotYetImplemented()
rstandard.cglm <- function (model, ...) .NotYetImplemented()
```

```
rstudent.cglm <- function (model, ...) .NotYetImplemented()
update.cglm <- function (object, ...) .NotYetImplemented()
deviance.cglm <- function (object, ...) .NotYetImplemented()
```

The `plot()` method, for example, will include residual plots like those shown above.

Most of the other methods listed here are similarly straightforward to implement. (The `deviance` method is a notable exception: its definition and documentation need special care, in light of the lack of a deviance function that is minimized by the generalized Wedderburn quasi-likelihood equations.)

Chapter 6

Conclusion

6.1 Summary of the Thesis Results

The main aim of this thesis is that of developing a model for the influence of explanatory variables on continuous compositional response variables. In Chapter 2, a multivariate logit model which may be used to model compositional data even if zeros are present in the data has been developed. This multivariate logit model generalizes an elegant method that was suggested previously by Wedderburn (1974) for the analysis of leaf blotch data in the special case of $J = 2$. In contrast to the logratio modeling approach devised by Aitchison (1982, 1986), the multivariate logit model used under the generalized Wedderburn approach models $E(Y_{ij})$ directly. The estimation of the parameters in this model is carried out using the technique of generalized estimating equations. This technique relies on the specification of a working correlation/variance-covariance structure. An appropriate working variance-covariance structure, $\phi\mathbb{V}_{\mathbf{p}_i,\boldsymbol{\Omega},\mathbb{W}}$, which caters for the variability arising in compositional data has been achieved (see equation (2.61)). The form of this working variance-covariance structure is based on the first order Taylor series approximation to the variance-covariance matrix of the composition $\mathbf{Y}_i$ where the components in $\mathbf{Y}_i$ are obtained as a result of taking the closure operation (1.1) on the corresponding latent variables $\dot{Y}_{ij}$. These latent variables are assumed to have a mean-variance relationship that generalizes the notion of a constant coefficient of variation.

As per Liang and Zeger (1986), the GEE estimator that is used to estimate the parameters of the multivariate logit model is actually a GLS estimator that has been shown to be invariant to the values of the correlation and dispersion parameters in the working variance-covariance structure (see Section 2.4.4). The invariance property of these estimators is analogous to the well-established invariance property of GLS estimators in multivariate linear regression (e.g. Mardia et al., 1979, p. 173). So in solving the estimating equations to obtain estimates of the model coefficients in the multivariate logit model, the 'independence' working variance-covariance structure $\mathbb{V}_{\mathbf{p}_i,\mathbb{I}_J,\mathbb{I}_J}$, defined in (2.63), may be used instead of $\phi\mathbb{V}_{\mathbf{p}_i,\boldsymbol{\Omega},\mathbb{W}}$ for computational simplicity. Also, due to the invariance prop-

erty and the fact that the estimating equations used under the generalized Wedderburn method (2.54), are linear and unbiased, the GEE estimator achieves full efficiency (has minimal generalized variance) across a wide class of potential dispersion and correlation matrices for the compositional response variables (McCullagh, 1983). Just as with any other GEE estimator, the GEE estimator used in the generalized Wedderburn method is also asymptotically unbiased and consistent, provided that the marginal mean model specification is correct.

In an analogy to the multivariate regression case, where the GLS estimator is free of the variance-covariance parameters but the variance-covariance matrix of the estimator is not, in Section 3.7.2.2 it has been shown that the asymptotic variance-covariance matrix of the GEE estimator is also a function of the correlation and dispersion parameters. A model-based variance estimator which takes into account of the variability of compositional data has thus been developed (see Section 2.8.2). Since the true variance-covariance matrix of $\mathbf{Y}_i$ is typically unknown, the model-based variance estimator proposed in this thesis 'borrows strength across subjects' (Liang and Zeger, 1986) by means of the 'squared Pearson residual' matrix for each i, to estimate the true variance-covariance matrix of $\mathbf{Y}_i$. The idea of pooling information from all the subjects to estimate $\text{Var}(\mathbf{Y}_i)$ has also be used by Pan (2001b). Our proposed variance estimator is in fact in the same form as Pan's estimator and so it also inherits the properties of the estimator proposed by Pan (2001b). By using the assumptions that the assumed variance-covariance structure is correct and that there is a common correlation across all cases i, the estimator proposed by Pan (2001b) lacks a degree of robustness when compared to the robust variance estimator proposed by Liang and Zeger (1986), but Pan's estimator has been proved to achieve greater efficiency asymptotically. Simulation studies carried out by Pan (2001b) suggest that greater efficiency for Pan's model-based estimator holds even for small sample sizes. Our model-based variance estimator has also been shown to be a direct generalization of the variance estimator used by Wedderburn (1974): the general method devised to estimate standard errors under the generalized Wedderburn approach agrees exactly with that of Wedderburn (1974) for the special case $J = 2$ (see Section 2.8.2).

In this thesis, the model-based variance estimator developed for use with the generalized Wedderburn approach has been studied empirically in a variety of situations. The model-based and robust variance estimates obtained when the Arctic Lake dataset and the Foraminiferal dataset were analyzed in Sections 3.9 and 3.10 respectively, were not appreciably different. In the small simulation study in which 10^5 compositional datasets were generated through a Dirichlet model (see Section 4.4), the robust variance estimator fared better at estimating the variance of the GEE estimator than the model-based estimator. The fact that the Liang and Zeger (1986) robust estimator achieved better estimates than the model-based estimator was however not surprising, since the Dirichlet model from which the data has been generated, is based on a different variance-covariance structure than the one considered in the generalized Wedderburn approach.

A comparison between the model-based and robust variance estimator was also carried out in another small simulation study in which 10^5 compositional datasets were generated through multivariate lognormally distributed $\dot{\mathbf{Y}}$ using three different sample sizes (60, 180 and 600), two different sets of coefficients of variation ($(5\%, 5\%, 20\%)$ and $(30\%, 30\%, 60\%)$), and three different correlations (independence, 0.3 and 0.7) (see Section 3.8.2). In this case, the model-based estimator being proposed in this thesis showed overall superiority over the robust estimator. The robust variance estimator underestimated the sample variance throughout. Consequently, the coverage probabilities obtained using the robust variance estimator were smaller than the nominal 95%. The coverage probabilities of the model-based estimator were much better, very close to 95% across all conditions. The results obtained for the model-based estimator in this study agree with the theoretical and simulation results obtained by Pan (2001b).

The simulation study with compositional data generated using the multivariate lognormal distribution was also carried out to compare the efficiency of the GEE estimator with that of the ML estimator used in the regression model devised by Aitchison (1982, 1986). In Chapter 3, the generalized Wedderburn method and Aitchison's regression method have been shown to share a number of formal similarities including the form of the estimator used, the form of the variance-covariance matrix of their respective estimator, the centering operation in the respective residuals. The two methods are also both related to the multiplicative model defined for the latent variables $\dot{Y}_{ij}$. Aitchison's regression method has in fact been shown to be an additive model which is obtained as a result of taking the logarithm of the multiplicative model (see Section 3.3). However, since under the generalized Wedderburn method, the mean-model specified for compositional $\mathbf{Y}$ is not the actual mean of $\mathbf{Y}$ but a first order Taylor series approximation to it, the generalized Wedderburn method and Aitchison's regression method estimate different mean models (see Section 3.4). In spite of this, efficiency comparisons between the estimators used in the two different models could still be undertaken if the truth from which the simulation datasets were generated was considered to have no dependence on the explanatory variables. In this way, the generalized Wedderburn method and Aitchison's method both had to estimate model coefficients whose true value is 0. What stood out in this simulation study is the fact that the variances obtained using the generalized Wedderburn method resulted to be either very similar to those achieved using Aitchison's approach or even slightly smaller. The GEE estimator was found to achieve the same or even slightly better efficiency than the ML estimator across all sample sizes and across all conditions considered. This behaviour might seem quite surprising, particularly when a sample as large as 600 is used and knowing that the ML estimator is well renowned for being a uniformly minimum variance unbiased estimator asymptotically. A theoretical explanation of why the GEE estimator may obtain better efficiency than the ML estimator has been presented in Section 3.8.2. This explanation is based on a comparison of the variance-covariance matrix of the ML estimator used in Aitchison's method with an estimate of the asymptotic variance-covariance matrix of the GEE estimator used in the generalized Wedderburn method.

Further to the just mentioned, a part of this thesis was also dedicated to devise various measures that may be used for model criticism of the generalized Wedderburn method. In a typical GEE analysis, Pan's Quasi Information Criterion (Pan, 2001a) is used for both variable selection and working correlation matrix selection. Due to the invariance property of the GEE estimator used in the generalized Wedderburn approach, the steps that are undertaken to check the quality of fit of the model do not involve choosing the best working correlation structure. With regards to variable selection, despite its popularity and ease of implementation, Pan's Quasi Information Criterion may not be used under the generalized Wedderburn approach. Pan's QIC (see equation (2.83)) relies on the specification of a log quasi-likelihood function. On using the estimating function U_{js}, defined in (2.55), for $j \neq j^{'}$ and $j, j^{'} = 1, \ldots, J-1$, it may be shown that $\partial U_{js}/\partial \gamma_{j^{'}k} \neq \partial U_{j^{'}k}/\partial \gamma_{js}$. Since the matrix of derivatives is not symmetric, the quasi log-likelihood function under the generalized Wedderburn method is thus not uniquely defined.

Testing whether model coefficients should be removed from the multivariate logit model or not may be carried out using the working Wald statistic (Rotnizky and Jewell, 1990) in conjunction with the model-based variance estimator proposed in this thesis (see Section 3.6.2). An alternative test statistic which tests whether model coefficients should be introduced in the multivariate logit model, and which also makes use of the model-based variance estimator proposed in this thesis, has also been presented. This test statistic, also due to Rotnizky and Jewell (1990), is the working score statistic. For detail on the two test statistics see Section 2.9.2.1.

Pearson residuals for the generalized Wedderburn method and a distance measure $\Delta\left(\mathbb{Y}, \widehat{\mathbb{P}}\right)$ which is based on Pearson residuals have also been developed in this thesis. A correspondence between the method devised by Wedderburn (1974) for $J = 2$ and the generalized Wedderburn method could once again be seen in the fact that the Pearson residuals obtained under the generalized Wedderburn are the same as the working residuals. Besides, the Pearson chi-square statistic computed using the Pearson residuals obtained under the generalized Wedderburn approach with $J = 2$ is exactly proportional to the Pearson chi-square statistic obtained under the model used by Wedderburn (1974).

The directed distance measure $\Delta\left(\mathbb{Y}, \widehat{\mathbb{P}}\right)$ may be used to check how close are the fitted values to the compositional response data. Aitchison (1992, p. 374) listed a number of criteria that should be satisfied by a distance measure used with compositions (see Pg 75). The distance measure $\Delta\left(\mathbb{Y}, \widehat{\mathbb{P}}\right)$ satisfies nearly all of these criteria, except for interchangeability of compositions and subcompositional dominance. The lack of interchangeability of the distance measure $\Delta\left(\mathbb{Y}, \widehat{\mathbb{P}}\right)$ should not however be viewed as a problem as to check the goodness of fit of a model it makes sense to check how far is a vector of fitted values $\widehat{\mathbf{p}}_i$ from the vector of compositions $\mathbf{Y}_i$ but not vice versa. The requirement for subcompositional dominance might be deemed too strict for use with the generalized Wedderburn approach. The parameter estimates that are obtained from analyzing a full composition under the generalized Wedderburn approach are not in general the same as those obtained

when a subcomposition is analyzed. However, the model assumptions that are used to analyze subcompositions are consistent with those used to analyze a full composition. For example, for some reference component J, the logits $\log\left(E\left(Y_{ij}\right)/E\left(Y_{iJ}\right)\right)$ are all modeled as $\mathbf{x}_i^{'}\boldsymbol{\gamma}$ if either a full composition or a subcomposition is analyzed.

6.2 Further Work

- Investigate small sample bias-adjustment of the GEE estimator used to estimate the model coefficients of the multivariate logit model.
  In the simulation study carried out with compositional data being obtained as a result of performing the closure operation on multivariate lognormally distributed $\dot{\mathbf{Y}}$, some significant downward bias was achieved by the GEE estimator when the generalized Wedderburn method has been used with samples of size 60 in conjunction with high coefficients of variation (30%, 60%, 60%) and a correlation of 0.3 or 0.7. In the simulation study where samples of size 39 have been generated through a Dirichlet model it could also be noticed that the GEE estimator exhibited some bias. The property of asymptotic unbiasedness of GEE estimators is well known (Liang and Zeger, 1986). Limited literature is however available on performance of GEE estimators when the sample size used is small. The bias obtained by the GEE estimator in the two simulation studies, seems to suggest that the bias-adjusted estimators used in Paul et al. (2013) and Paul and Zhang (2014) might be worth investigating when using the generalized Wedderburn approach with small sample sizes.

- Perform a simulation study to compare the efficiency of the GEE estimator under the generalized Wedderburn approach with the ML estimator used to estimate the model coefficients of the Dirichlet regression model when the variance-covariance structure underlying the Dirichlet model is more complicated than a homogeneous variance model.

- Compare the performance of the multivariate logit model with other recently proposed strategies available for modeling compositional data, such as the Kent regression model of Scealy and Welsh (2011), the multivariate simplex model of Zhang (2013) and the α-regression method of Tsagris (2015).

- Study the performance of the multivariate logit model in the presence of a high percentage of zeros in the data.

- Consider extensions of the generalized Wedderburn method to ordered compositional response and to compositions with a hierarchical structure.

- Complete and polish the *cglm* package and its documentation, for publication on CRAN as a contributed package for the R language.

Appendix A

Deriving the Model-Based Estimator $\widehat{\text{Var}\,(\widehat{\gamma})}_M$ for $J = 2$ (see Section 2.8.1)

The model-based estimator $\widehat{\text{Var}\,(\widehat{\gamma})}_M$ for $J = 2$ is given by

$$\widehat{\text{Var}\,(\widehat{\gamma})}_M = \left[\frac{1}{n-(p+1)}\sum_{i=1}^{n}\frac{(Y_{i1}-\hat{p}_{i1})^2}{\hat{p}_{i1}^2\,(1-\hat{p}_{i1})^2}\right]\left(\mathbb{X}'\mathbb{X}\right)^{-1}. \tag{A.1}$$

The derivation of the above result starts by considering the form of $\widehat{\boldsymbol{\Sigma}}_i^*$.

The estimator $\widehat{\boldsymbol{\Sigma}}_i^*$ for $J = 2$ is given by

$$\begin{aligned}
\widehat{\boldsymbol{\Sigma}}_i^* &= \frac{1}{4}\begin{pmatrix}1 & -1\\ -1 & 1\end{pmatrix}\begin{pmatrix}\frac{1}{\hat{p}_{i1}} & 0\\ 0 & \frac{1}{\hat{p}_{i2}}\end{pmatrix}\begin{pmatrix}Y_{i1}-\hat{p}_{i1}\\ Y_{i2}-\hat{p}_{i2}\end{pmatrix}\begin{pmatrix}Y_{i1}-\hat{p}_{i1}, & Y_{i2}-\hat{p}_{i2}\end{pmatrix}\begin{pmatrix}\frac{1}{\hat{p}_{i1}} & 0\\ 0 & \frac{1}{\hat{p}_{i2}}\end{pmatrix}\begin{pmatrix}1 & -1\\ -1 & 1\end{pmatrix}\\
&= \frac{1}{4}\begin{pmatrix}\frac{Y_{i1}}{\hat{p}_{i1}}-\frac{Y_{i2}}{\hat{p}_{i2}}\\ -\left(\frac{Y_{i1}}{\hat{p}_{i1}}-\frac{Y_{i2}}{\hat{p}_{i2}}\right)\end{pmatrix}\begin{pmatrix}\frac{Y_{i1}}{\hat{p}_{i1}}-\frac{Y_{i2}}{\hat{p}_{i2}}, & -\left(\frac{Y_{i1}}{\hat{p}_{i1}}-\frac{Y_{i2}}{\hat{p}_{i2}}\right)\end{pmatrix}\\
&= \frac{1}{4}\begin{pmatrix}\frac{(Y_{i1}-\hat{p}_{i1})^2}{\hat{p}_{i1}^2(1-\hat{p}_{i1})^2} & -\left(\frac{(Y_{i1}-\hat{p}_{i1})^2}{\hat{p}_{i1}^2(1-\hat{p}_{i1})^2}\right)\\ -\left(\frac{(Y_{i1}-\hat{p}_{i1})^2}{\hat{p}_{i1}^2(1-\hat{p}_{i1})^2}\right) & \frac{(Y_{i1}-\hat{p}_{i1})^2}{\hat{p}_{i1}^2(1-\hat{p}_{i1})^2}\end{pmatrix}\\
&= \frac{1}{4}\frac{(Y_{i1}-\hat{p}_{i1})^2}{\hat{p}_{i1}^2\,(1-\hat{p}_{i1})^2}\begin{pmatrix}1 & -1\\ -1 & 1\end{pmatrix},
\end{aligned}$$

since

$$\frac{Y_{i1}}{p_{i1}}-\frac{Y_{i2}}{p_{i2}} = \frac{Y_{i1}}{p_{i1}}-\frac{1-Y_{i1}}{1-p_{i1}} = \frac{Y_{i1}-p_{i1}}{p_{i1}\,(1-p_{i1})}.$$

Then

$$\widehat{\boldsymbol{\Sigma}}^* = \frac{1}{n-(p+1)}\sum_{i=1}^{n}\widehat{\boldsymbol{\Sigma}}_i^* = \frac{1}{4\,(n-(p+1))}\left[\sum_{i=1}^{n}\frac{(Y_{i1}-\hat{p}_{i1})^2}{\hat{p}_{i1}^2\,(1-\hat{p}_{i1})^2}\right]\begin{pmatrix}1 & -1\\ -1 & 1\end{pmatrix}. \tag{A.2}$$

The substitution of (A.2) in (2.74), leads to

$$\widehat{\hat{\phi}\mathbb{V}_{\mathbf{p}_{i'},\mathbf{\Omega},\mathbb{W}}} = \begin{pmatrix} 1 & -1 \\ -1 & 1 \end{pmatrix} \left[\frac{\hat{p}_{i'1}^2 \left(1-\hat{p}_{i'1}\right)^2}{4\left(n-(p+1)\right)} \left[\sum_{i=1}^{n} \frac{\left(Y_{i1}-\hat{p}_{i1}\right)^2}{\hat{p}_{i1}^2 \left(1-\hat{p}_{i1}\right)^2} \right] \begin{pmatrix} 1 & -1 \\ -1 & 1 \end{pmatrix} \right] \begin{pmatrix} 1 & -1 \\ -1 & 1 \end{pmatrix} \tag{A.3}$$

$$= \frac{\hat{p}_{i'1}^2 \left(1-\hat{p}_{i'1}\right)^2}{4\left(n-(p+1)\right)} \left[\sum_{i=1}^{n} \frac{\left(Y_{i1}-\hat{p}_{i1}\right)^2}{\hat{p}_{i1}^2 \left(1-\hat{p}_{i1}\right)^2} \begin{pmatrix} 4 & -4 \\ -4 & 4 \end{pmatrix} \right]$$

$$= \frac{\hat{p}_{i'1}^2 \left(1-\hat{p}_{i'1}\right)^2}{n-(p+1)} \left[\sum_{i=1}^{n} \frac{\left(Y_{i1}-\hat{p}_{i1}\right)^2}{\hat{p}_{i1}^2 \left(1-\hat{p}_{i1}\right)^2} \begin{pmatrix} 1 & -1 \\ -1 & 1 \end{pmatrix} \right].$$

Using the estimator of (2.49) and substituting (A.3) in (2.67) gives the required expression:

$$\widehat{\mathrm{Var}\left(\widehat{\boldsymbol{\gamma}}\right)}_M = \left(\sum_{i'=1}^{n} \hat{p}_{i'1}^2 \left(1-\hat{p}_{i'1}\right)^2 \mathbb{X}_{i'}' \begin{pmatrix} 1 & -1 \\ -1 & 1 \end{pmatrix} \right. \tag{A.4}$$

$$\left. \times \left[\frac{\hat{p}_{i'1}^2 \left(1-\hat{p}_{i'1}\right)^2}{n-(p+1)} \sum_{i=1}^{n} \frac{\left(Y_{i1}-\hat{p}_{i1}\right)^2}{\hat{p}_{i1}^2 \left(1-\hat{p}_{i1}\right)^2} \begin{pmatrix} 1 & -1 \\ -1 & 1 \end{pmatrix} \right]^{-} \begin{pmatrix} 1 & -1 \\ -1 & 1 \end{pmatrix} \mathbb{X}_{i'} \right)^{-1}$$

$$= \frac{1}{n-(p+1)} \sum_{i=1}^{n} \frac{\left(Y_{i1}-\hat{p}_{i1}\right)^2}{\hat{p}_{i1}^2 \left(1-\hat{p}_{i1}\right)^2} \left(\sum_{i'=1}^{n} \mathbb{X}_{i'}' \begin{pmatrix} 1 & -1 \\ -1 & 1 \end{pmatrix} \mathbb{X}_{i'} \right)^{-1}$$

$$= \left[\frac{1}{n-(p+1)} \sum_{i=1}^{n} \frac{\left(Y_{i1}-\hat{p}_{i1}\right)^2}{\hat{p}_{i1}^2 \left(1-\hat{p}_{i1}\right)^2} \right] \left(\mathbb{X}'\mathbb{X} \right)^{-1},$$

since

$$\begin{pmatrix} 1 & -1 \\ -1 & 1 \end{pmatrix} \begin{pmatrix} 1 & -1 \\ -1 & 1 \end{pmatrix}^{-} \begin{pmatrix} 1 & -1 \\ -1 & 1 \end{pmatrix} = \begin{pmatrix} 1 & -1 \\ -1 & 1 \end{pmatrix}.$$

Appendix B

Proof to show that the Matrix of Derivatives is not Symmetric (see Section 2.9.2.2)

It is required to show that for $j \neq j^{''}$

$$\frac{\partial U_{js}}{\partial \gamma_{j^{''}k}} \neq \frac{\partial U_{j^{''}k}}{\partial \gamma_{js}}.$$

Proof. From (2.55),

$$\begin{aligned} U_{js} &= \sum_{i=1}^{n} \left(\frac{Y_{ij}}{p_{ij}} - \frac{1}{J} \sum_{j^{'}=1}^{J} \frac{Y_{ij^{'}}}{p_{ij^{'}}} \right) x_{is} \\ &= \sum_{i=1}^{n} \left[\left(Y_{ij} \exp\left(-\mathbf{x}_i^{'} \boldsymbol{\gamma}_j\right) - \frac{1}{J} \sum_{j^{'}=1}^{J} Y_{ij^{'}} \exp\left(-\mathbf{x}_i^{'} \boldsymbol{\gamma}_{j^{'}}\right) \right) \sum_{j^{'''}=1}^{J} \exp\left(\mathbf{x}_i^{'} \boldsymbol{\gamma}_{j^{'''}}\right) \right] x_{is} \end{aligned}$$

so,

$$\begin{aligned} \frac{\partial U_{js}}{\partial \gamma_{j^{''}k}} &= \sum_{i=1}^{n} \left[\left(Y_{ij} \exp\left(-\mathbf{x}_i^{'} \boldsymbol{\gamma}_j\right) - \frac{1}{J} \sum_{j^{'}=1}^{J} Y_{ij^{'}} \exp\left(-\mathbf{x}_i^{'} \boldsymbol{\gamma}_{j^{'}}\right) \right) \exp\left(\mathbf{x}_i^{'} \boldsymbol{\gamma}_{j^{''}}\right) x_{ik} \right] x_{is} \\ &\quad + \sum_{i=1}^{n} \left[-\frac{1}{J} Y_{ij^{''}} \exp\left(-\mathbf{x}_i^{'} \boldsymbol{\gamma}_{j^{''}}\right) (-x_{ik}) \sum_{j^{'''}=1}^{J} \exp\left(\mathbf{x}_i^{'} \boldsymbol{\gamma}_{j^{'''}}\right) \right] x_{is} \\ &= \sum_{i=1}^{n} \left[\frac{Y_{ij}}{\exp\left(\mathbf{x}_i^{'}\left(\boldsymbol{\gamma}_j - \boldsymbol{\gamma}_{j^{''}}\right)\right)} - \frac{1}{J} \left(\sum_{j^{'}=1}^{J} \frac{Y_{ij^{'}}}{\exp\left(\mathbf{x}_i^{'}\left(\boldsymbol{\gamma}_{j^{'}} - \boldsymbol{\gamma}_{j^{''}}\right)\right)} - \frac{Y_{ij^{''}}}{p_{ij^{''}}} \right) \right] x_{ik} x_{is}. \end{aligned}$$

Similarly,

$$
\begin{aligned}
U_{j''k} &= \sum_{i=1}^{n} \left(\frac{Y_{ij''}}{p_{ij''}} - \frac{1}{J} \sum_{j'=1}^{J} \frac{Y_{ij'}}{p_{ij'}} \right) x_{ik} \\
&= \sum_{i=1}^{n} \left[\left(Y_{ij''} \exp\left(-\mathbf{x}_i^{'} \boldsymbol{\gamma}_{j''}\right) - \frac{1}{J} \sum_{j'=1}^{J} Y_{ij'} \exp\left(-\mathbf{x}_i^{'} \boldsymbol{\gamma}_{j'}\right) \right) \sum_{j'''=1}^{J} \exp\left(\mathbf{x}_i^{'} \boldsymbol{\gamma}_{j'''}\right) \right] x_{ik}
\end{aligned}
$$

$$
\begin{aligned}
\frac{\partial U_{j''k}}{\partial \gamma_{js}} &= \sum_{i=1}^{n} \left[\left(Y_{ij''} \exp\left(-\mathbf{x}_i^{'} \boldsymbol{\gamma}_{j''}\right) - \frac{1}{J} \sum_{j'=1}^{J} Y_{ij'} \exp\left(-\mathbf{x}_i^{'} \boldsymbol{\gamma}_{j'}\right) \right) \exp\left(\mathbf{x}_i^{'} \boldsymbol{\gamma}_{j}\right) x_{is} \right] x_{ik} \\
&\quad + \sum_{i=1}^{n} \left[-\frac{1}{J} Y_{ij} \exp\left(-\mathbf{x}_i^{'} \boldsymbol{\gamma}_{j}\right) (-x_{is}) \sum_{j'''=1}^{J} \exp\left(\mathbf{x}_i^{'} \boldsymbol{\gamma}_{j'''}\right) \right] x_{ik} \\
&= \sum_{i=1}^{n} \left[\frac{Y_{ij''}}{\exp\left(\mathbf{x}_i^{'} \left(\boldsymbol{\gamma}_{j''} - \boldsymbol{\gamma}_{j}\right)\right)} - \frac{1}{J} \left(\sum_{j'=1}^{J} \frac{Y_{ij'}}{\exp\left(\mathbf{x}_i^{'} \left(\boldsymbol{\gamma}_{j'} - \boldsymbol{\gamma}_{j}\right)\right)} - \frac{Y_{ij}}{p_{ij}} \right) \right] x_{is} x_{ik},
\end{aligned}
$$

which is not equal to $\partial U_{js} / \partial \gamma_{j''k}$ as required.

□

Appendix C

Proof to show Equality of Distance Measures (see Section 3.6.1)

It is required to show that

$$\sum_{j''=1}^{J} \underbrace{\sum_{j=1}^{J}}_{j<j''} \sum_{i=1}^{n} \left[\log\left(\frac{Y_{ij}}{\hat{p}^*_{ij}}\right) - \log\left(\frac{Y_{ij''}}{\hat{p}^*_{ij''}}\right) \right]^2$$
$$= J \sum_{j=1}^{J} \sum_{i=1}^{n} \left[\log\left(\frac{Y_{ij}}{\hat{p}^*_{ij}}\right) - \frac{1}{J} \sum_{j'=1}^{J} \log\left(\frac{Y_{ij'}}{\hat{p}^*_{ij'}}\right) \right]^2 .$$

Proof. From Section 3.6.1,

$$R^*_{ij} = \log\left(\frac{Y_{ij}}{\hat{p}^*_{ij}}\right) - \frac{1}{J} \sum_{j'=1}^{J} \log\left(\frac{Y_{ij'}}{\hat{p}^*_{ij'}}\right),$$

and $\sum_{j=1}^{J} R^*_{ij} = 0$. Also,

$$
\begin{aligned}
&\sum_{j''=1}^{J} \underbrace{\sum_{j=1}^{J}}_{j<j''} \left[\log\left(\frac{Y_{ij}}{\hat{p}^*_{ij}}\right) - \log\left(\frac{Y_{ij''}}{\hat{p}^*_{ij''}}\right) \right]^2 \\
&= \sum_{j''=1}^{J} \underbrace{\sum_{j=1}^{J}}_{j<j''} \left[\log\left(\frac{Y_{ij}}{\hat{p}^*_{ij}}\right) - \frac{1}{J}\sum_{j'=1}^{J} \log\left(\frac{Y_{ij'}}{\hat{p}^*_{ij'}}\right) \right. \\
&\quad \left. - \log\left(\frac{Y_{ij''}}{\hat{p}^*_{ij''}}\right) + \frac{1}{J}\sum_{j'=1}^{J} \log\left(\frac{Y_{ij'}}{\hat{p}^*_{ij'}}\right) \right]^2 \\
&= \sum_{j''=1}^{J} \underbrace{\sum_{j=1}^{J}}_{j<j''} \left(R^*_{ij} - R^*_{ij''}\right)^2.
\end{aligned}
$$

Then

$$
\begin{aligned}
\sum_{j=1}^{J} \underbrace{\sum_{j''=1}^{J}}_{j<j'} \left(R^*_{ij} - R^*_{ij''}\right)^2 &= \frac{1}{2}\sum_{j=1}^{J} \underbrace{\sum_{j''=1}^{J}}_{j\neq j''} \left(R^*_{ij} - R^*_{ij''}\right)^2 \\
&= \frac{1}{2}\sum_{j=1}^{J}\sum_{j''=1}^{J} \left(R^*_{ij} - R^*_{ij''}\right)^2 \\
&= \frac{1}{2}\sum_{j=1}^{J}\sum_{j''=1}^{J} \left(R^{*2}_{ij} + R^{*2}_{ij''} - 2R^*_{ij''}R^*_{ij}\right) \\
&= \frac{1}{2}\left(J\sum_{j''=1}^{J} R^{*2}_{ij} + J\sum_{j=1}^{J} R^{*2}_{ij''}\right) \\
&= J\sum_{j=1}^{J} R^{*2}_{ij},
\end{aligned}
$$

as required. □

Appendix D

Proof to show that the Row Vectors of $\mathbb{F}\left(\mathbb{P}_i - \mathbf{p}_i\mathbf{p}_i^{'}\right)$ Sum to Zero and are Linearly Independent (see Section 3.7.2.2)

Proof.

Part 1. Let the column vectors of $\left(\mathbb{P}_i - \mathbf{p}_i\mathbf{p}_i^{'}\right)$ be represented by $\mathbf{C}_{(j)}$, $(j = 1, \ldots, J)$, where $\mathbf{C}_{(j)} = (C_{1j}, \ldots, C_{Jj})^{'}$.Then

$$\mathbb{F}\left(\mathbb{P}_i - \mathbf{p}_i\mathbf{p}_i^{'}\right) = \left(\mathbb{F}\mathbf{C}_{(1)}, \ldots, \mathbb{F}\mathbf{C}_{(J)}\right)$$

and since interest lies in the sum of the rows of $\mathbb{F}\left(\mathbb{P}_i - \mathbf{p}_i\mathbf{p}_i^{'}\right)$, we turn our attention to $\mathbb{F}\sum_{j=1}^{J}\mathbf{C}_{(j)}$. Now, for any given row $j^{'}$, $\left(j^{'} = 1, \ldots, J\right)$,

$$\sum_{j=1}^{J} C_{j^{'}j} = p_{ij^{'}} - p_{ij^{'}}\sum_{j=1}^{J} p_{ij} = 0$$

since $\sum_{j=1}^{J} p_{ij} = 1$.

It therefore follows that $\sum_{j=1}^{J}\mathbf{C}_{(j)} = \mathbf{0}$ leading to $\mathbb{F}\sum_{j=1}^{J}\mathbf{C}_{(j)} = \mathbf{0}$, showing that the row sums of $\mathbb{F}\left(\mathbb{P}_i - \mathbf{p}_i\mathbf{p}_i^{'}\right)$ are all equal to zero.

Part 2. Let the row vectors of $\mathbb{F}$ be represented by $\boldsymbol{F}_{j^*}^{'}$, $(j^* = 1, \ldots, J-1)$, where $\boldsymbol{F}_{j^*} = (F_{j^*1}, \ldots, F_{j^*J})^{'}$. Then, the elements in $\mathbb{F}\left(\mathbb{P}_i - \mathbf{p}_i\mathbf{p}_i^{'}\right)$ are given by

$$\left[\mathbb{F}\left(\mathbb{P}_i - \mathbf{p}_i\mathbf{p}_i^{'}\right)\right]_{j^*j} = \boldsymbol{F}_{j^*}^{'}\mathbf{C}_{(j)},$$

and the j^{*th} row vector of $\mathbb{F}\left(\mathbb{P}_i - \mathbf{p}_i\mathbf{p}_i^{'}\right)$ is given by $\left(\boldsymbol{F}_{j^*}^{'}\mathbf{C}_{(1)}, \ldots, \boldsymbol{F}_{j^*}^{'}\mathbf{C}_{(J)}\right)$.

Now, to show that the row vectors of $\mathbb{F}\left(\mathbb{P}_i - \mathbf{p}_i\mathbf{p}_i{}'\right)$ are linearly independent, we need to show that any linear combination of the row vectors is equal to 0 only if the coefficients are equal to 0.

For any given coefficients $u_1, \ldots, u_{J-1}$, $u_{j^*} \in \mathbb{R}$, the linear combinations of interest are:

$$u_1 \begin{pmatrix} \boldsymbol{F}_1'\mathbf{C}_{(1)} \\ \vdots \\ \boldsymbol{F}_1'\mathbf{C}_{(J)} \end{pmatrix} + \cdots + u_{J-1} \begin{pmatrix} \boldsymbol{F}_{J-1}'\mathbf{C}_{(1)} \\ \vdots \\ \boldsymbol{F}_{J-1}'\mathbf{C}_{(J)} \end{pmatrix} = \mathbf{0}. \tag{D.1}$$

The linear combinations (D.1) may be reexpressed as

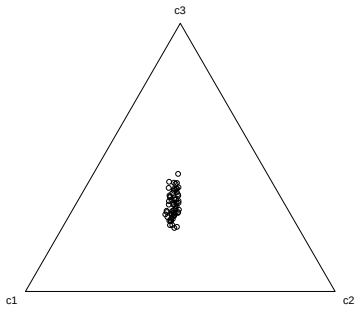
$$\begin{pmatrix} \mathbf{C}_{(1)}\left(u_1\boldsymbol{F}_1' + \cdots + u_{J-1}\boldsymbol{F}_{J-1}'\right) \\ \vdots \\ \mathbf{C}_{(J)}\left(u_1\boldsymbol{F}_1' + \cdots + u_{J-1}\boldsymbol{F}_{J-1}'\right) \end{pmatrix} = \mathbf{0}. \tag{D.2}$$

Under the assumption that the proportions making up the elements $C_{j'j}$ are not equal to zero and given that the rows of $\mathbb{F}$ are linearly independent, it is only possible for the linear combinations in (D.2) to be equal to zero if the coefficients $u_1, \ldots, u_{J-1}$ are equal to 0, showing that the $J-1$ rows of $\mathbb{F}\left(\mathbb{P}_i - \mathbf{p}_i\mathbf{p}_i{}'\right)$ are indeed linearly independent.

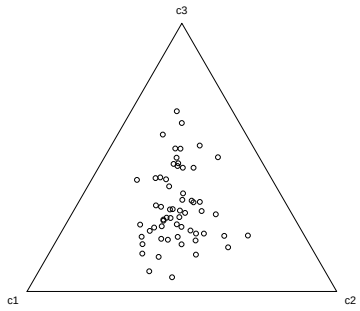
□

Appendix E

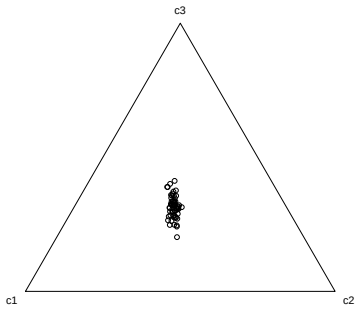
Ternary Diagrams of Simulated Datasets for Combinations of Sample Size, Correlation and Coefficients of Variation (see Section 3.8)



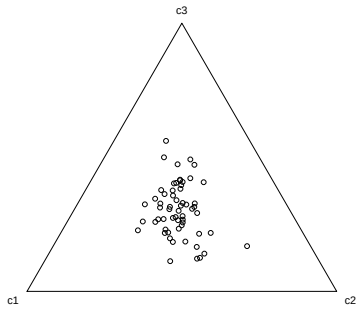
(a) Ternary Diagram for the First Generated Sample of size 60 assuming independence and coefficients of variation $(5\%, 5\%, 20\%)$



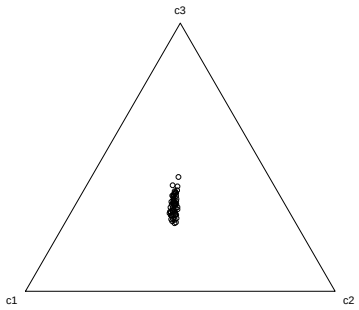
(b) Ternary Diagram for the First Generated Sample of size 60 assuming independence and coefficients of variation $(30\%, 30\%, 60\%)$



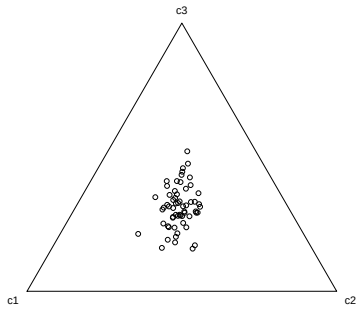
(c) Ternary Diagram for the First Generated Sample of size 60 assuming correlation 0.3 and coefficients of variation $(5\%, 5\%, 20\%)$



(d) Ternary Diagram for the First Generated Sample of size 60 assuming correlation 0.3 and coefficients of variation $(30\%, 30\%, 60\%)$

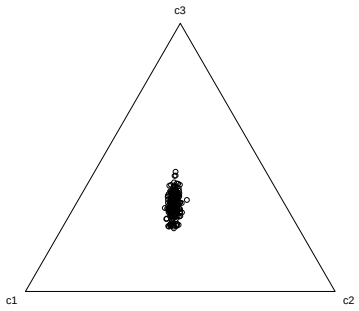


(e) Ternary Diagram for the First Generated Sample of size 60 assuming correlation 0.7 and coefficients of variation $(5\%, 5\%, 20\%)$

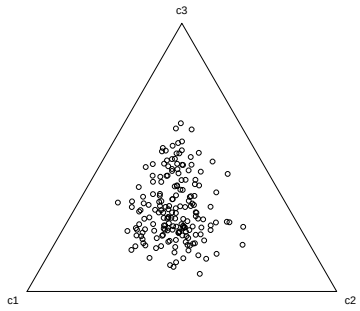


(f) Ternary Diagram for the First Generated Sample of size 60 assuming correlation 0.7 and coefficients of variation $(30\%, 30\%, 60\%)$

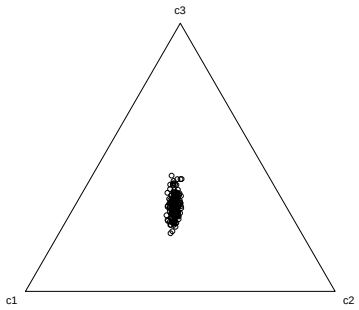
Figure E.1: Ternary Diagrams of the First Generated Sample of Size 60 assuming different Correlations and Coefficients of Variation



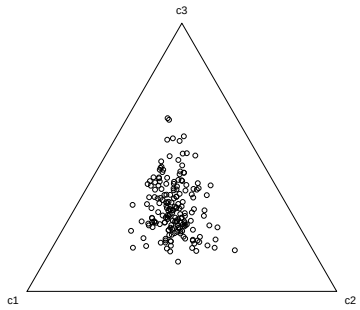
(a) Ternary Diagram for the First Generated Sample of size 180 assuming independence and coefficients of variation $(5\%, 5\%, 20\%)$



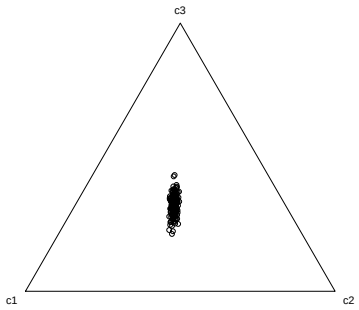
(b) Ternary Diagram for the First Generated Sample of size 180 assuming independence and coefficients of variation $(30\%, 30\%, 60\%)$



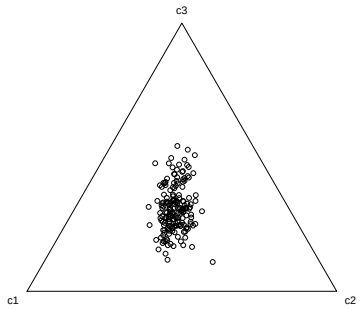
(c) Ternary Diagram for the First Generated Sample of size 180 assuming correlation 0.3 and coefficients of variation $(5\%, 5\%, 20\%)$



(d) Ternary Diagram for the First Generated Sample of size 180 assuming correlation 0.3 and coefficients of variation $(30\%, 30\%, 60\%)$

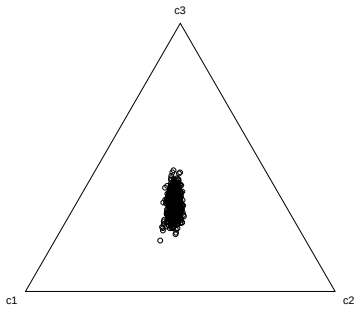


(e) Ternary Diagram for the First Generated Sample of size 180 assuming correlation 0.7 and coefficients of variation $(5\%, 5\%, 20\%)$

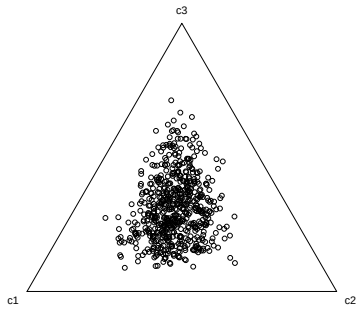


(f) Ternary Diagram for the First Generated Sample of size 180 assuming correlation 0.7 and coefficients of variation $(30\%, 30\%, 60\%)$

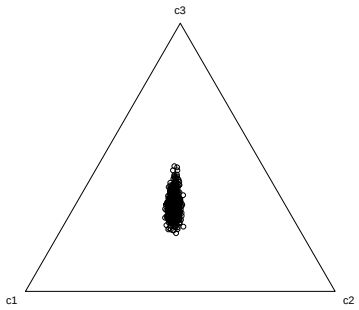
Figure E.2: Ternary Diagrams of the First Generated Sample of Size 180 assuming different Correlations and Coefficients of Variation



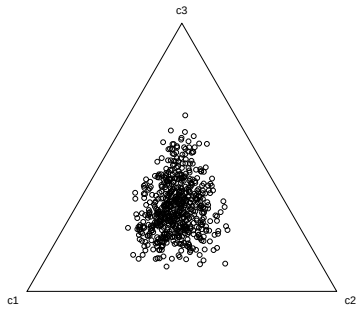
(a) Ternary Diagram for the First Generated Sample of size 600 assuming independence and coefficients of variation $(5\%, 5\%, 20\%)$



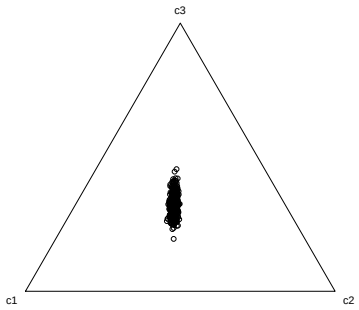
(b) Ternary Diagram for the First Generated Sample of size 600 assuming independence and coefficients of variation $(30\%, 30\%, 60\%)$



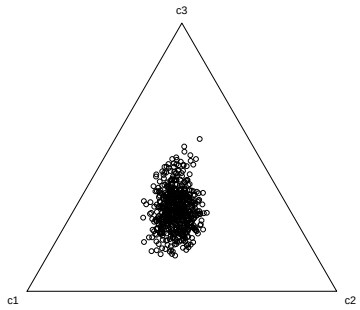
(c) Ternary Diagram for the First Generated Sample of size 600 assuming correlation 0.3 and coefficients of variation $(5\%, 5\%, 20\%)$



(d) Ternary Diagram for the First Generated Sample of size 600 assuming correlation 0.3 and coefficients of variation $(30\%, 30\%, 60\%)$



(e) Ternary Diagram for the First Generated Sample of size 600 assuming correlation 0.7 and coefficients of variation $(5\%, 5\%, 20\%)$



(f) Ternary Diagram for the First Generated Sample of size 600 assuming correlation 0.7 and coefficients of variation $(30\%, 30\%, 60\%)$

Figure E.3: Ternary Diagrams of the First Generated Sample of Size 600 assuming different Correlations and Coefficients of Variation

Appendix F

Scatter Plots of Generalized Wedderburn Estimates versus Aitchison Estimates for γ_{11} and γ_{21} for Combinations of Sample Size, Correlation and Coefficients of Variation (see Section 3.8.2)

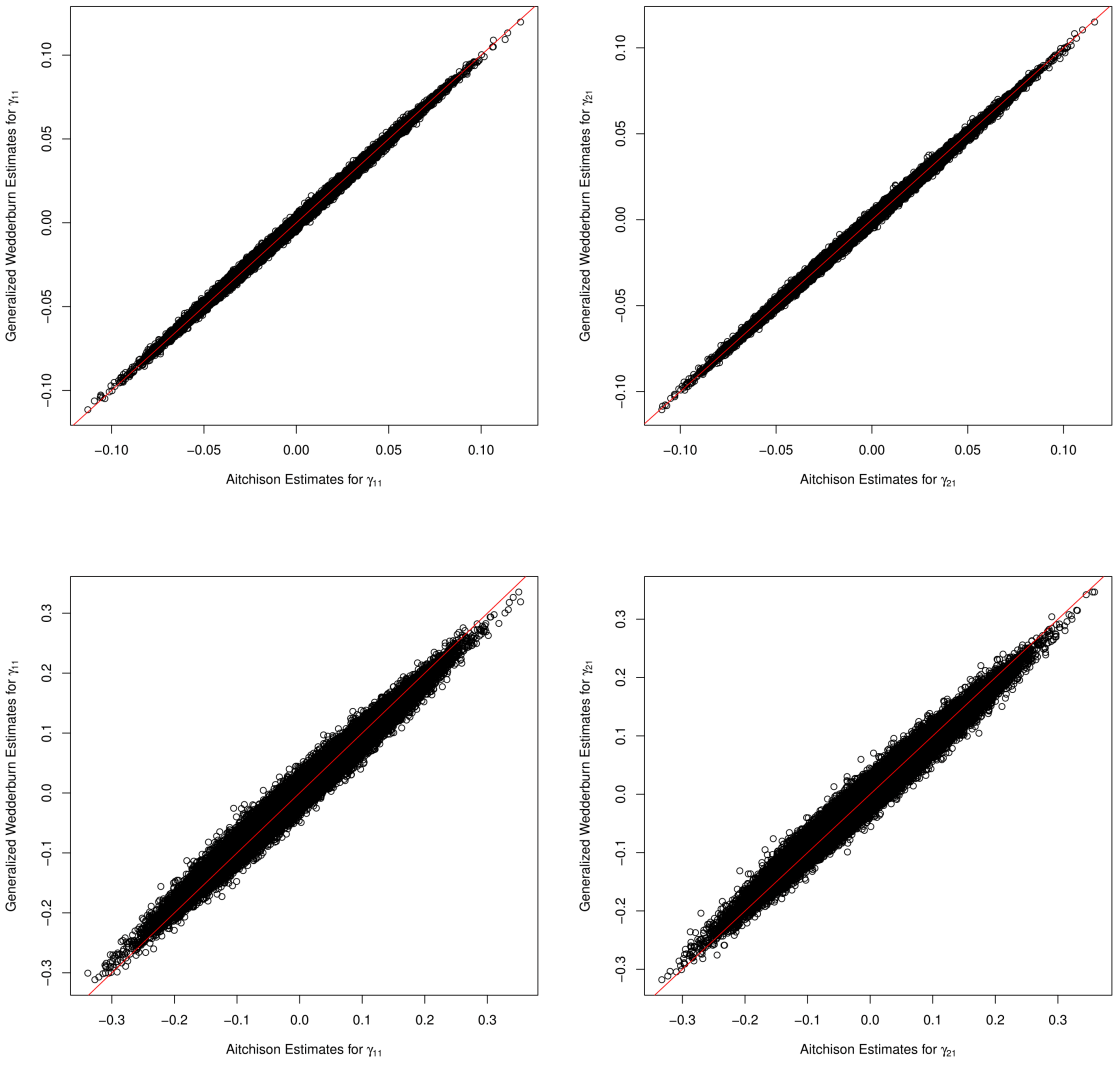


Figure F.1: Scatter Plots of Generalized Wedderburn versus Aitchison Estimates using a simulation size of 10^5, samples of size 60, assuming independence, and coefficients of variation $(5\%, 5\%, 20\%)$ and $(30\%, 30\%, 60\%)$ in the upper and lower panes respectively

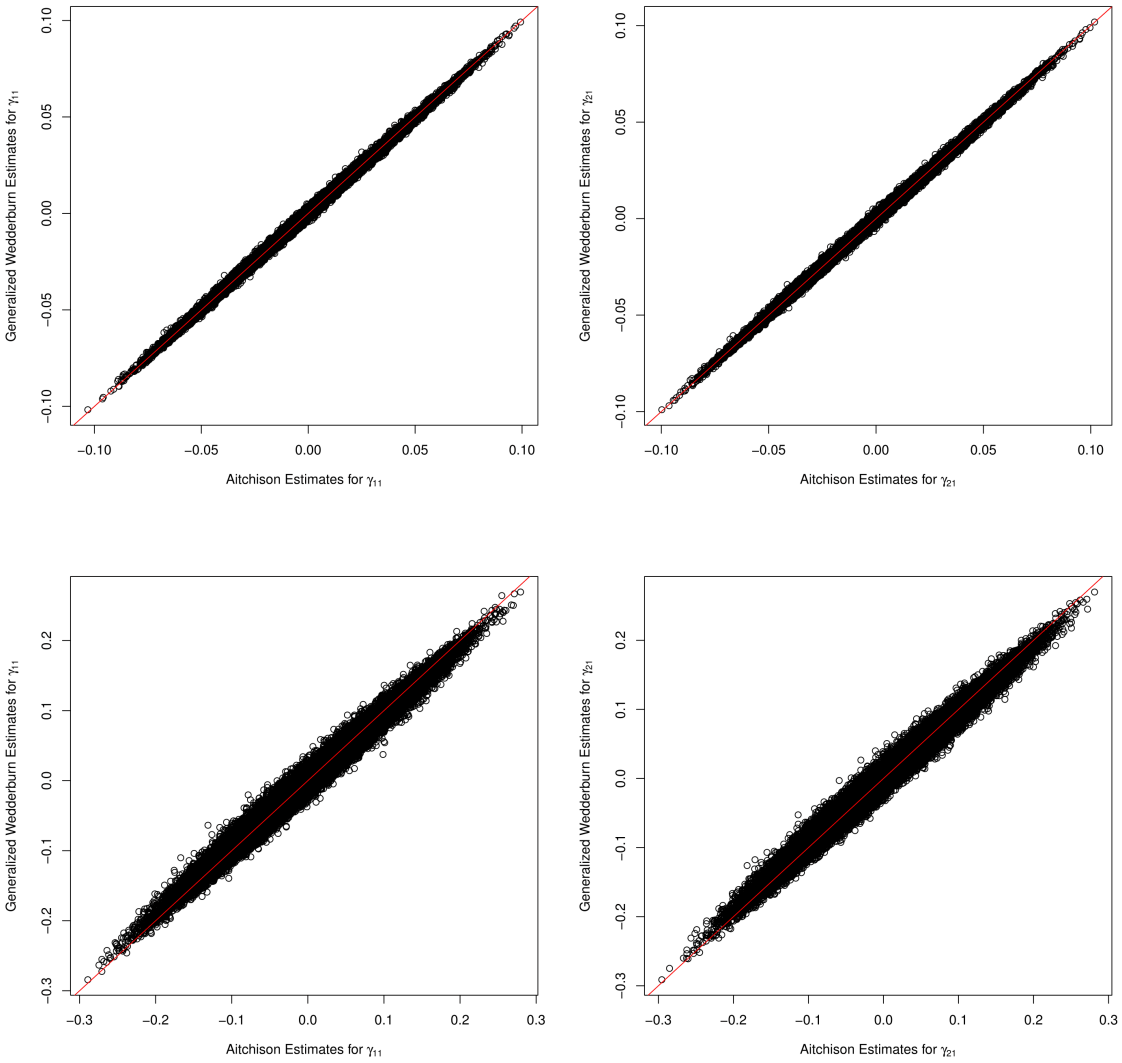


Figure F.2: Scatter Plots of Generalized Wedderburn versus Aitchison Estimates using a simulation size of 10^5, samples of size 60, assuming correlation 0.3, and coefficients of variation $(5\%, 5\%, 20\%)$ and $(30\%, 30\%, 60\%)$ in the upper and lower panes respectively

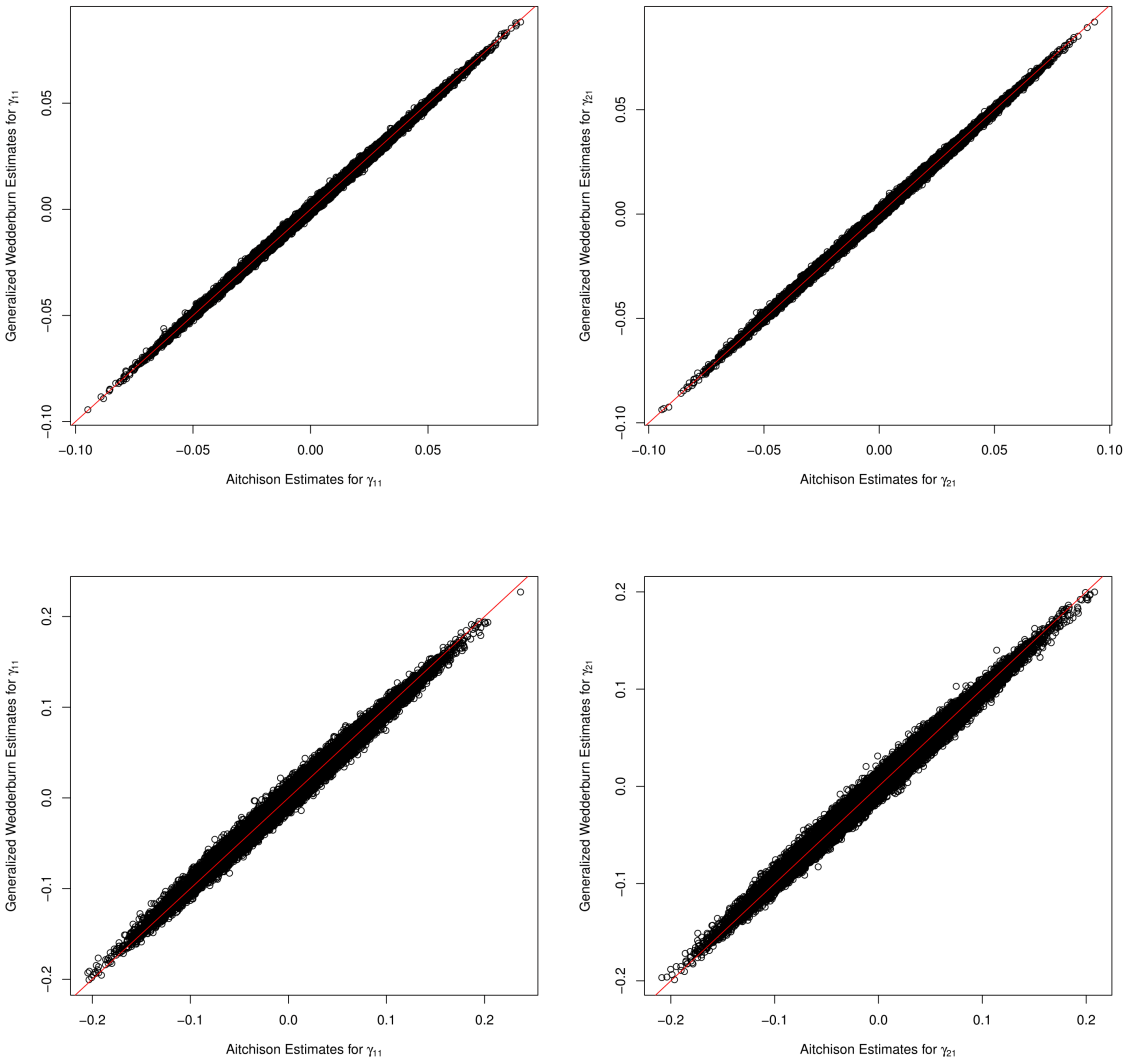


Figure F.3: Scatter Plots of Generalized Wedderburn versus Aitchison Estimates using a simulation size of 10^5, samples of size 60, assuming correlation 0.7, and coefficients of variation $(5\%, 5\%, 20\%)$ and $(30\%, 30\%, 60\%)$ in the upper and lower panes respectively

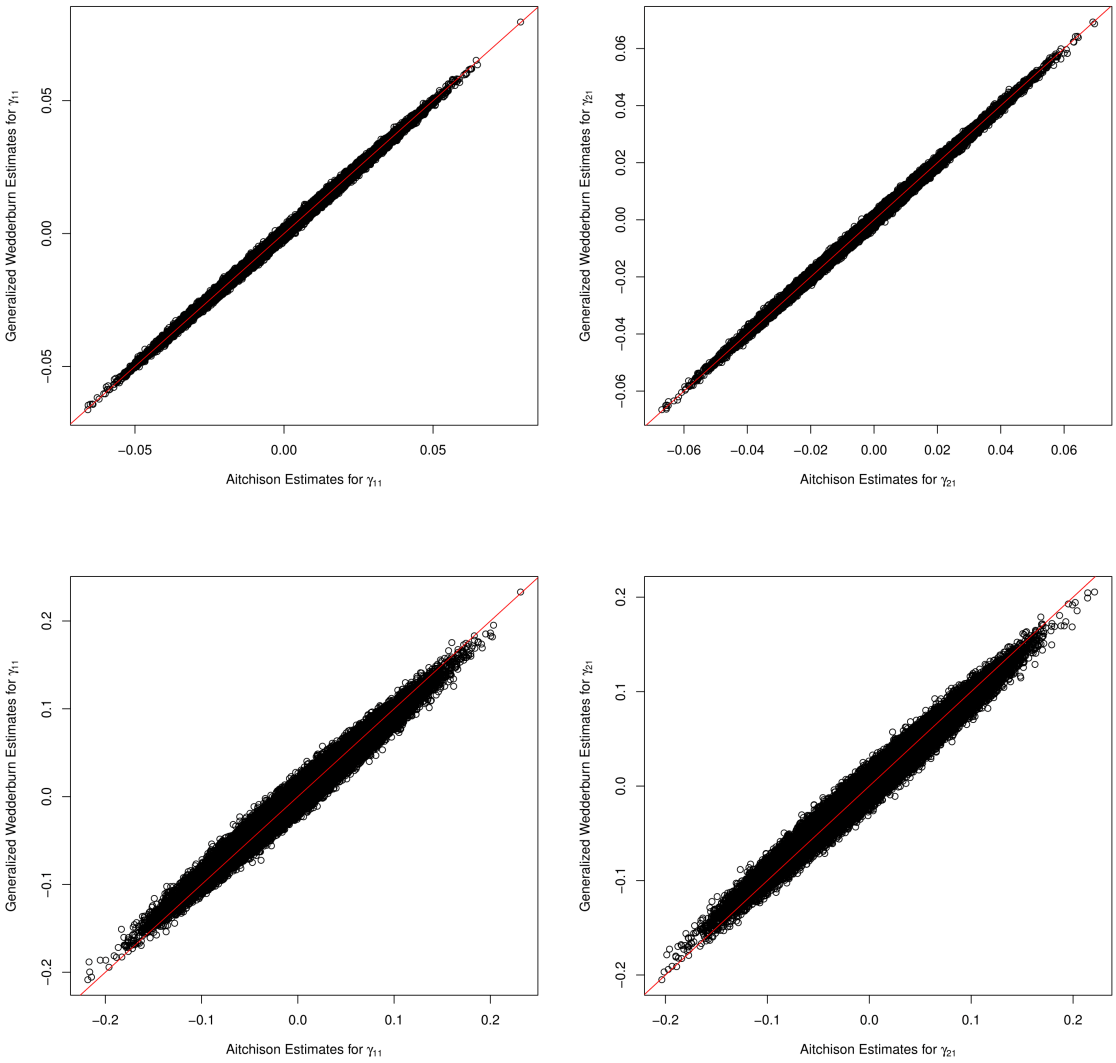


Figure F.4: Scatter Plots of Generalized Wedderburn versus Aitchison Estimates using a simulation size of 10^5, samples of size 180, assuming independence, and coefficients of variation $(5\%, 5\%, 20\%)$ and $(30\%, 30\%, 60\%)$ in the upper and lower panes respectively

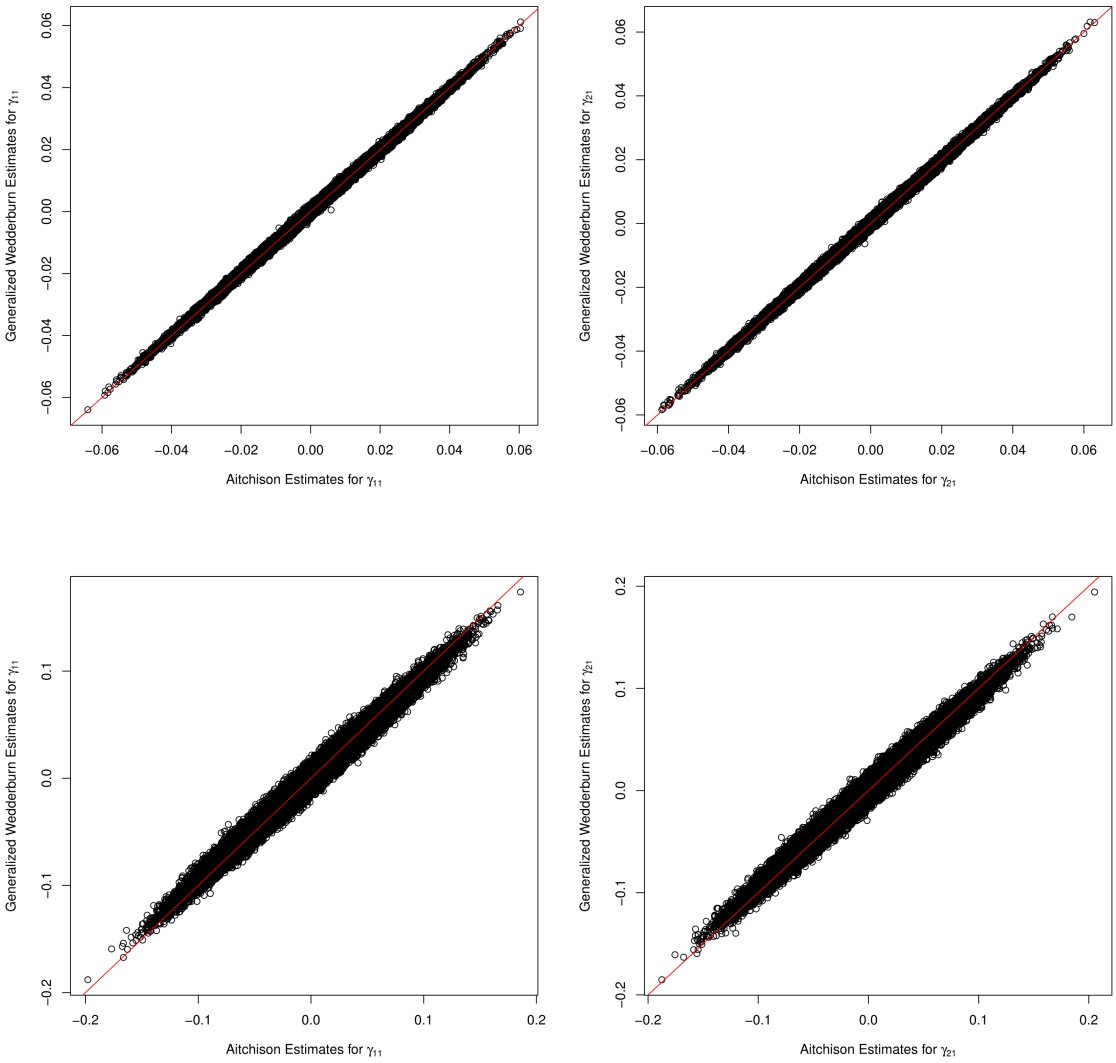


Figure F.5: Scatter Plots of Generalized Wedderburn versus Aitchison Estimates using a simulation size of 10^5, samples of size 180, assuming correlation 0.3, and coefficients of variation $(5\%, 5\%, 20\%)$ and $(30\%, 30\%, 60\%)$ in the upper and lower panes respectively

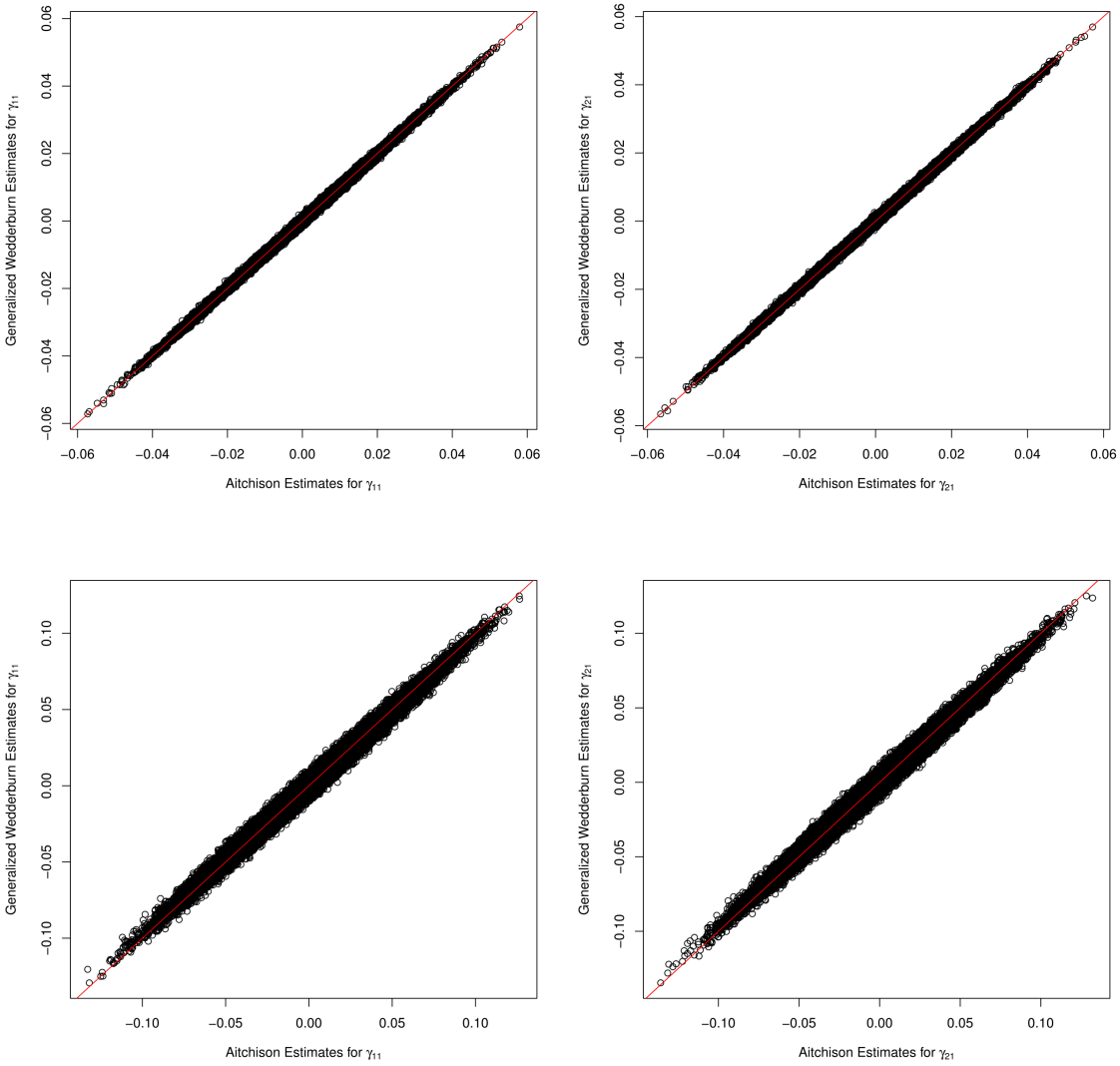


Figure F.6: Scatter Plots of Generalized Wedderburn versus Aitchison Estimates using a simulation size of 10^5, samples of size 180, assuming correlation 0.7, and coefficients of variation $(5\%, 5\%, 20\%)$ and $(30\%, 30\%, 60\%)$ in the upper and lower panes respectively

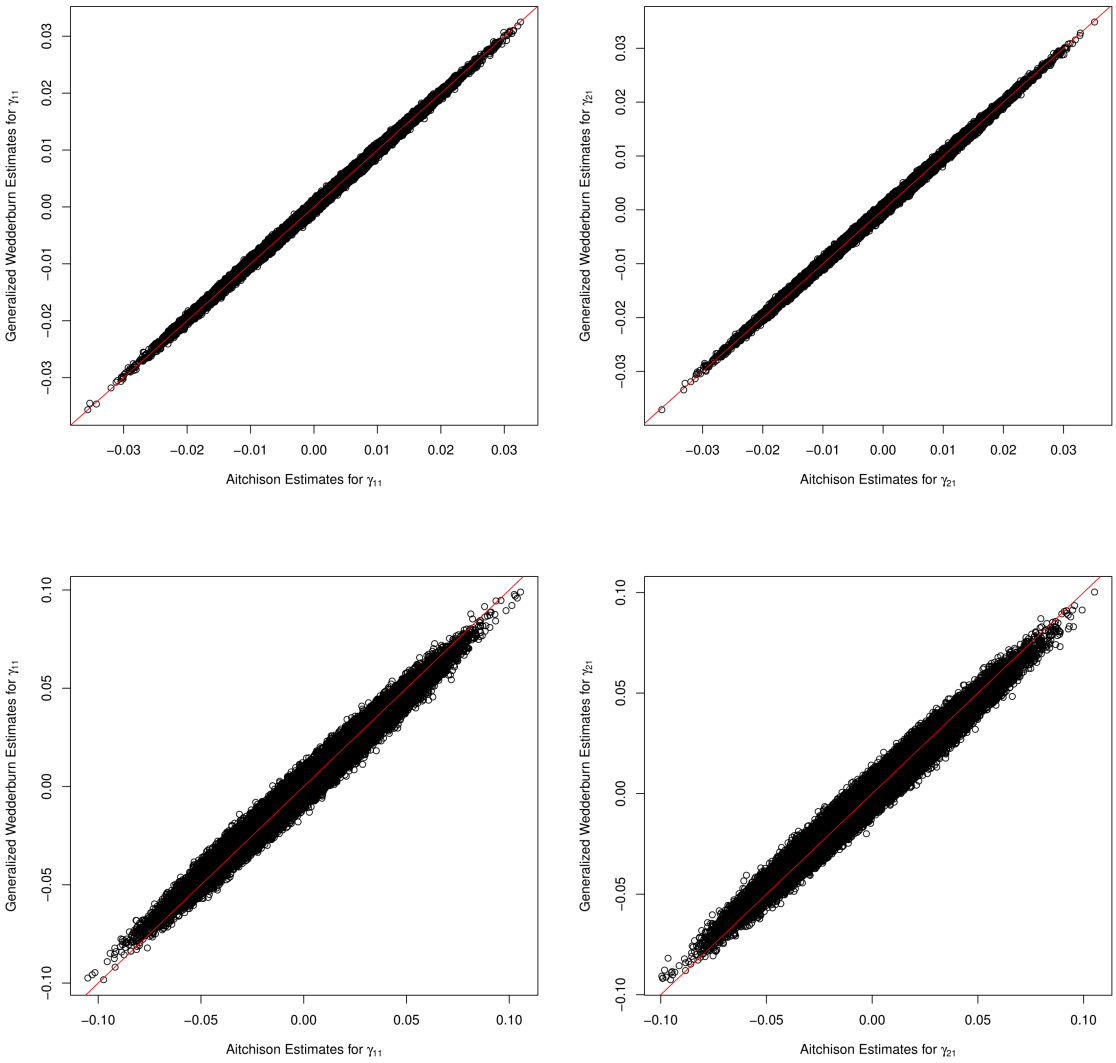


Figure F.7: Scatter Plots of Generalized Wedderburn versus Aitchison Estimates using a simulation size of 10^5, samples of size 600, assuming independence, and coefficients of variation $(5\%, 5\%, 20\%)$ and $(30\%, 30\%, 60\%)$ in the upper and lower panes respectively

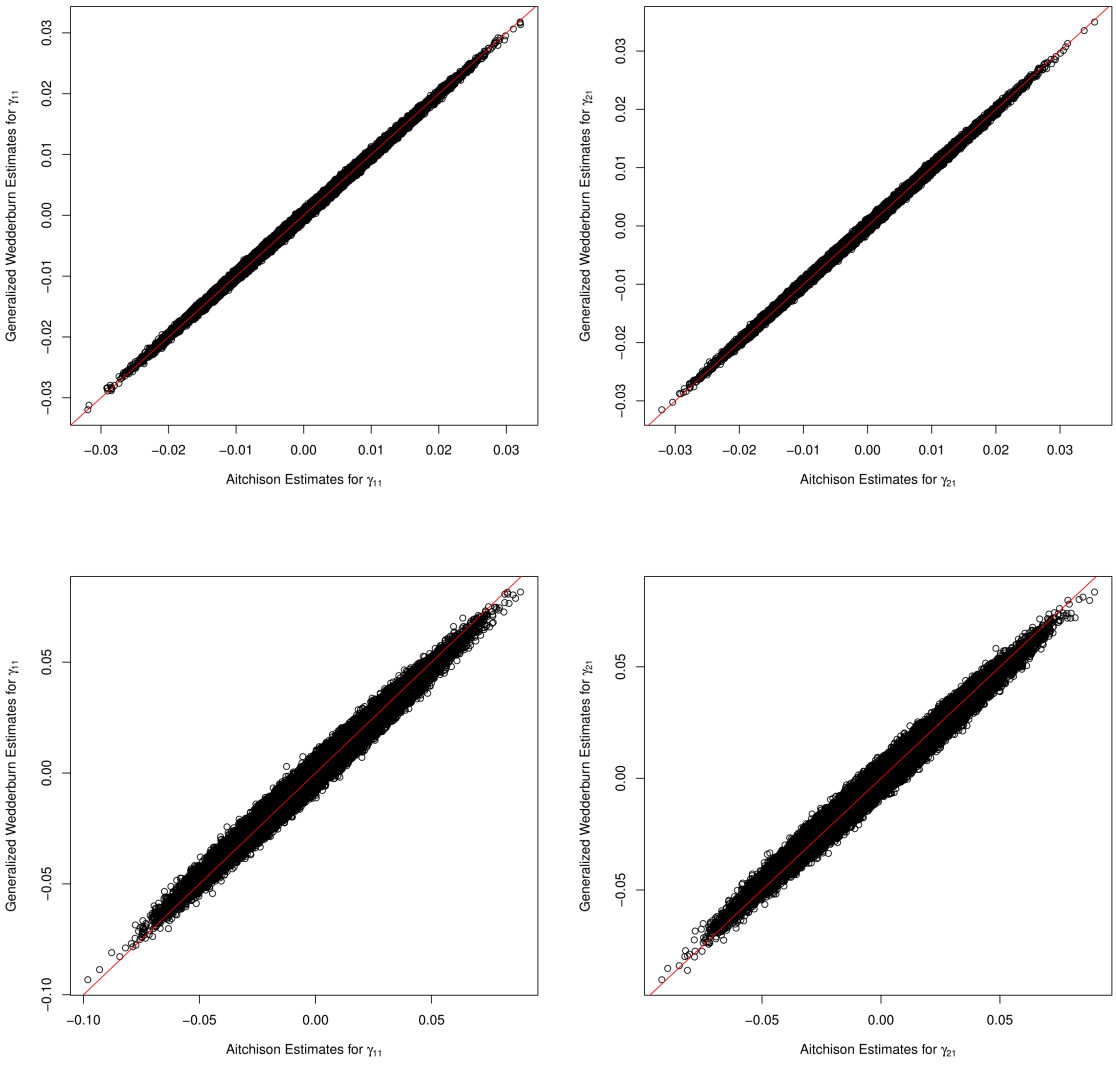


Figure F.8: Scatter Plots of Generalized Wedderburn versus Aitchison Estimates using a simulation size of 10^5, samples of size 600, assuming correlation 0.3, and coefficients of variation $(5\%, 5\%, 20\%)$ and $(30\%, 30\%, 60\%)$ in the upper and lower panes respectively

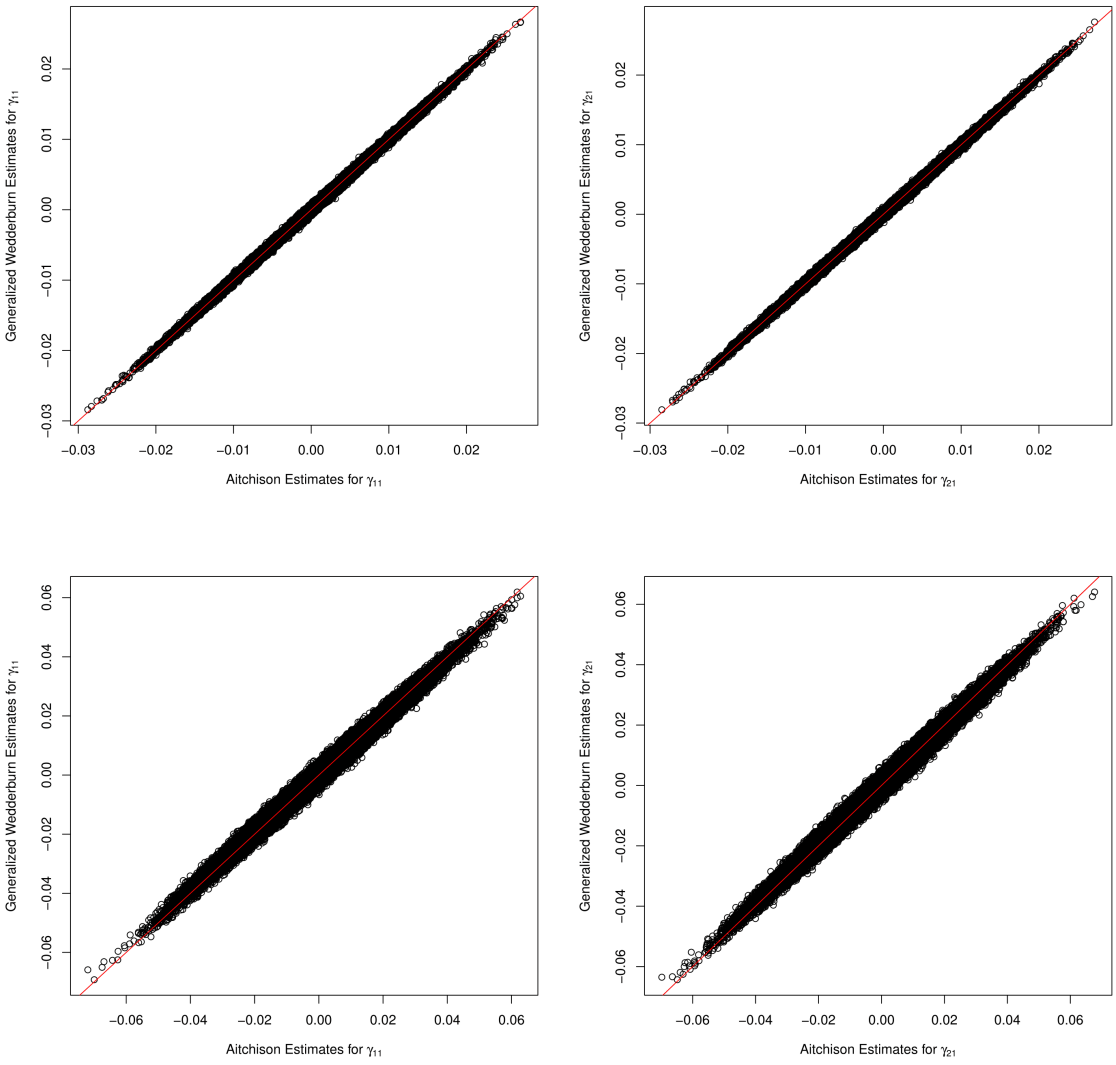


Figure F.9: Scatter Plots of Generalized Wedderburn versus Aitchison Estimates using a simulation size of 10^5, samples of size 600, assuming correlation 0.7, and coefficients of variation $(5\%, 5\%, 20\%)$ and $(30\%, 30\%, 60\%)$ in the upper and lower panes respectively

Bibliography

Agnew, J., Balduzzi, P., and Sundén, A. (1995). Portfolio choice and trading in a large 401 (k) plan. *American Economic Review*, 93:11–23.

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

Aitchison, J. (1984). The statistical analysis of geochemical compositions. *Journal of the International Association for Mathematical Geology*, 16:531–564.

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

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

Aitchison, J. (2003a). Compositional data analysis: Where are we and where should we be heading? In *Thió-Henestrosa, S. Martín-Fernández, J. A. (eds) Compositional Data Analysis Workshop, Girona, Spain*. University of Girona, electronic publication http://ima.udg.es/Activitats/CoDaWork03/.

Aitchison, J. (2003b). *The Statistical Analysis of Compositional Data*. The Blackburn Press.

Aitchison, J. (2008). The single principal of compositional data analysis, continuing fallacies, confusions and misunderstandings and some suggested remedies. In *CODAWORK'08, Girona, Spain*.

Aitchison, J. and Barceló-Vidal, C. (2001). Reply to letter to the editor by S. Rehder and U. Ziehr on 'Logratio analysis and compositional distance' by Aitchison, J. and Barceló-Vidal, c. and Martín-Fernández, J. A. and Pawlowsky-Glahn, V. *Mathematical Geology*, 33:849–860.

Aitchison, J. and Brown, J. A. C. (1957). *The Lognormal Distribution*. Cambridge University Press.

Aitchison, J. and Kay, J. W. (2003). Possible solutions of some essential zero problems in compositional data analysis. In *Thió-Henestrosa, S. Martín-Fernández, J. A. (eds)*

Compositional Data Analysis Workshop, Girona, Spain. University of Girona, electronic publication http://ima.udg.es/Activitats/CoDaWork03/.

Aitchison, J. and Shen, S. M. (1980). Logistic-normal distributions: Some properties and uses. *Biometrika*, 67:261–272.

Anderson, J. A. (1972). Separate sample logistic discrimination. *Biometrika*, 59:19–35.

Bacon-Shone, J. (2003). Modelling structural zeros in compositional data. In *Thió-Henestrosa, S. Martín-Fernández, J. A. (eds) Compositional Data Analysis Workshop, Girona, Spain*. University of Girona, electronic publication http://ima.udg.es/Activitats/CoDaWork03/.

Barclay, M. J. and Smith, C. W. (1995). The determinants of corporate leverage and dividend policies. *Journal of Applied Corporate Finance*, 7:4–19.

Barndorff-Nielsen, O. E. and Jørgensen, B. (1991). Some parametric models on the simplex. *Journal of Multivariate Analysis*, 39(1):106 – 116.

Boos, D. D. (1992). On generalized score tests. *The American Statistician*, 46:327–333.

Box, G. E. P. and Cox, D. R. (1964). An analysis of transformations. *Journal of the Royal Statistical Society, Series B, Methodological*, 26(2):211–252.

Butler, A. and Glasbey, C. (2008). A latent Gaussian model for compositional data. *Journal of the Royal Statistical Society, Series C*, 57:505–520.

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

Connor, J. R. and Mosimann, J. E. (1969). Concepts of independence for proportions with a generalization of the Dirichlet distribution. *Journal of American Statistical Association*, 64:194–206.

Cox, C. (1996). Nonlinear quasi-likelihood models: Applications to continuous proportions. *Computational Statistics & Data Analysis*, 21:449–461.

Cox, D. and Snell, E. J. (1968). A general definition of residuals. *Journal of the Royal Statistical Society, Series B, Methodological*, 30(2):248–275.

Darroch, J. N. (1969). Null correlation for proportions. *Mathematical Geology*, 1:221–227.

Darroch, J. N. and James, I. R. (1978). F-independence and null correlations of bounded-sum, positive variables. *Journal of the Royal Statistical Society, Series B, Methodological*, 36:467–483.

Darroch, J. N. and Ratcliff, D. (1970). Null correlation for proportions-ii. *Mathematical Geology*, 2:307–312.

Darroch, J. N. and Ratcliff, D. (1978). No-association of proportions. *Mathematical Geology*, 10:361–368.

Drum, M. and McCullagh, P. (1993). [Regression models for discrete longitudinal responses]: Comment. *Statistical Science*, 8(3):300–301.

Emrich, L. and Piedmonte, M. (1992). On some small sample properties of generalized estimating equation estimates for multivariate dichotomous outcomes. *Journal of Statistical Computation and Simulation*, 41:19–29.

Fay, M. and Graubard, B. (2001). Small-sample adjustments for Wald-type tests using sandwich estimators. *Biometrics*, 57:1198–1206.

Firth, D. (1987). On the efficiency of quasi-likelihood estimation. *Biometrika*, 74:233–245.

Firth, D. (1988). Multiplicative errors: Log-normal or gamma? *Journal of the Royal Statistical Society, Series B, Methodological*, 50:266–268.

Firth, D. (1993a). Bias reduction of maximum likelihood estimates. *Biometrika*, 80:27–38.

Firth, D. (1993b). Recent developments in quasi-likelihood methods. *Bulletin of the International Statistical Institute*, 55:341–358.

Fitzmaurice, G. (1995). A caveat concerning independence estimating equations with multivariate binary data. *Biometrics*, 51:309–317.

Gilchrist, R. (1982). An analysis of continuous proportions. In Caussinus, H., Ettinger, P., and Tomassone, R., editors, *COMPSTAT 1982 5th Symposium held at Toulouse 1982: Part I: Proceedings in Computational Statistics*, pages 236–241. Physica-Verlag HD, Heidelberg.

Gosho, M., Hamada, C., and Yoshimura, I. (2011). Modifications of QIC and CIC for selecting a working correlation structure in the generalized estimating equation method. *Japanese Journal of Biometrics*, 32:1–12.

Gosho, M., Sato, Y., and Takeuchi, H. (2014). Robust covariance estimator for small-sample adjustment in the generalized estimating equations: A simulation study. *Science Journal of Applied Mathematics and Statistics*, 2:20–25.

Gourieroux, C., Monfort, A., and Trognon, A. (1984). Pseudo maximum likelihood methods: Theory. *Econometrica*, 52:681–700.

Gueorguieva, R., Rosenheck, R., and Zelterman, D. (2008). Dirichlet component regression and its applications to psychiatric data. *Computational Statistics & Data Analysis*, 52(12):53445355.

Guo, X., Pan, W., Connett, J. E., Hannan, P. J., and French, S. A. (2005). Small-sample performance of the robust score test and its modifications in generalized estimating equations. *Statistics in Medicine*, 24(22):3479–3495.

Gupta, R. D. and Richards, D. S. P. (1987). Multivariate Liouville distributions. *Journal of Multivariate Analysis*, 23(2):233–256.

Gupta, R. D. and Richards, D. S. P. (1991). Multivariate Liouville distributions, ii. *Probability and Mathematical Statistics*, 12(2):291–309.

Gupta, R. D. and Richards, D. S. P. (1992). Multivariate Liouville distributions, iii. *Journal of Multivariate Analysis*, 43(1):29–57.

Gupta, R. D. and Richards, D. S. P. (1995). Multivariate Liouville distributions, iv. *Journal of Multivariate Analysis*, 54(1):1–17.

Hanfelt, J. J. and Liang, K. (1995). Approximate likelihood ratios for general estimating functions. *Biometrika*, 82(3):461–477.

Hin, L. and Wang, Y. (2009). Working-correlation-structure identification in generalized estimating equations. *Statistics in Medicine*, 27:642–658.

Huber, P. (1967). The behavior of maximum likelihood estimates under nonstandard conditions. In *Proceedings of the Fifth Berkeley Symposium on Mathematical Statistics and Probability, Volume 1: Statistics*, pages 221–233, Berkeley, California. University of California Press.

James, I. R. and Mosimann, J. E. (1980). A new characterization of the Dirichlet distribution through neutrality. *The Annals of Statistics*, 8:183–189.

Jørgensen, B. (1986). Some properties of exponential dispersion models. *Scandinavian Journal of Statistics*, 13(3):187–197.

Kauermann, G. and Carroll, R. (2001). A note on the efficiency of sandwich covariance matrix estimation. *Journal of the American Statistical Association*, 96:1387–1396.

Kent, J. T. (1982). The Fisher-Bingham distribution on the sphere. *Journal of the Royal Statistical Society, Series B, Methodological*, 44(1):71–80.

Kieschnick, R. and McCullough, B. D. (2003). Regression analysis of variates observed on (0,1): Percentages, proportions and fractions. *Statistical Modelling*, 3:193–213.

Kork, J. O. (1977). Examination of the Chayes-Kruskal procedure for testing correlations between proportions. *Mathematical Geology*, 9:543–562.

Kosmidis, I. and Firth, D. (2011). Multinomial logit bias reduction via the Poisson log-linear model. *Biometrika*, 98:755–759.

Kullback, S. (1997). *Information Theory and Statistics*. Dover Publications.

Lancaster, H. O. (1965). The Helmert matrices. *The American Mathematical Monthly*, 72(1):4–12.

Li, B. (1993). A deviance function for the quasi-likelihood method. *Biometrika*, 80:741–753.

Li, B. and McCullagh, P. (1994). Potential functions and conservative estimating functions. *The Annals of Statistics*, 22:340–356.

Liang, K. and Zeger, S. L. (1986). Longitudinal data analysis using generalized linear models. *Biometrika*, 73:13–22.

Lindley, D. V. (1964). The Bayesian analysis of contingency tables. *The Annals of Mathematical Statistics*, 35:1622–1643.

Maier, M. J. (2014). Dirichletreg: Dirichlet regression for compositional data in R. Technical Report 125, Institute for Statistics and Mathematics, Vienna University of Economics and Business.

Mancl, L. A. and DeRouen, T. A. (2001). A covariance estimator for gee with improved small-sample properties. *Biometrics*, 57(1):126–134.

Mardia, K. V. (1976). Discussion on 'The ordering of multivariate data' (by V. Barnett). *Journal of the Royal Statistical Society, Series A*, 139:346–347.

Mardia, K. V. and Jupp, P. E. (2000). *Directional Statistics*. Wiley.

Mardia, K. V., Kent, J. T., and Bibby, J. M. (1979). *Multivariate Analysis*. Academic Press.

Martín-Fernández, J. A., Barceló-Vidal, C., and Pawlowsky-Glahn, V. (2003). Dealing with zeros and missing values in compositional data sets using nonparametric imputation. *Mathematical Geology*, 35:253–278.

McAlister, D. (1879). The law of the geometric mean. *Proceedings of the Royal Society of London*, 29:367–376.

McCullagh, P. (1983). Quasi-likelihood functions. *Annals of Statistics*, 11:59–67.

McCullagh, P. (1990). Quasi-likelihood and estimating functions. In *Statistical Theory and Modeling: In Honour of Sir David Cox, FRS, Eds. D.V. Hinkley, N. Reid, E. J. Snell, pp 265-286.* Chapman and Hall/CRC.

McCullagh, P. and Nelder, J. A. (1989). *Generalized Linear Models, Second Edition*. Chapman and Hall/CRC.

McLeish, D. L. and Small, C. G. (1992). A projected likelihood function for semiparametric models. *Biometrika*, 79:93–102.

Meisch, A. T. (1969). The constant sum problem in geochemistry. In *Computer Applications in the Earth Sciences, (D. F. Merriam, ed.)*, pages 161–177.

Migliorati, S., Ongaro, A., and Monti, G. S. (2016). A structured Dirichlet mixture model for compositional data: inferential and applicative issues. *Statistics and Computing*, pages 1–21.

Ongaro, A. and Migliorati, S. (2013). A generalization of the Dirichlet distribution. *Journal of Multivariate Analysis*, 114:412–426.

Paik, M. C. (1992). Parametric variance function estimation for nonnormal repeated measurement data. *Biometrics*, 48:19–30.

Palarea-Albaladejo, J. and Martín-Fernández, J. A. (2008). A modified EM alr-algorithm for replacing rounded zeros in compositional data sets. *Computers & Geosciences*, 34:902–917.

Palarea-Albaladejo, J., Martín-Fernández, J. A., and Gómez-García, J. (2007). A parametric approach for dealing with compositional rounded zeros. *Mathematical Geology*, 39:625–645.

Palmgren, J. (1981). The Fisher information matrix for log linear models arguing conditionally on observed explanatory variable. *Biometrika*, 68:563–566.

Pan, W. (2001a). Akaike's information criterion in generalized estimating equations. *Biometrics*, 57:120–125.

Pan, W. (2001b). On the robust variance estimator in generalised estimating equations. *Biometrika*, 88(3):901–906.

Pan, W. and Wall, M. M. (2002). Small-sample adjustments in using the sandwich variance estimator in generalized estimating equations. *Statistics in Medicine*, 21:1429–1441.

Papke, L. E. and Wooldridge, J. M. (1996). Econometric methods for fractional response variables with an application to 401(k) plan participation rates. *Journal of Applied Econometrics*, 11:619–632.

Paul, S. and Zhang, X. (2014). Small sample gee estimation of regression parameters for longitudinal data. *Statistics in Medicine*, 33(22):3869–3881.

Paul, S., Zhang, X., and Xu, J. (2013). Estimation of regression parameters for binary longitudinal data using GEE: Review, extension and an application to environmental data. *Journal of Environmental Statistics*, 4(1).

Pawlowsky-Glahn, V. and Egozcue, J. J. (2006). Compositional data and their analysis: An introduction. *Geological Society, London, Special Publications*, 264:1–10.

Prentice, R. L. (1988). Correlated binary regression with covariates specific to each binary observation. *Biometrics*, 44:1033–1048.

R Core Team (2016). *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing, Vienna, Austria.

Rayens, S. R. and Srinivasan, C. (1994). Dependence properties of generalized Liouville distributions on the simplex. *Journal of the American Statistical Association*, 89(428):1465–1470.

Rehder, S. and Zier, U. (2001). Letters to the editor, Comment on 'logratio analysis and compositional distance' by Aitchison, J. and Barceló-Vidal, C. and Martín-Fernández, J. A. and Pawlowsky-Glahn, V. *Mathematical Geology*, 33:845–848.

Rotnizky, A. and Jewell, R. (1990). Hypothesis testing of regression parameters in semiparametric generalized linear models for cluster correlated data. *Biometrika*, 77:485–497.

Scealy, J. L. and Welsh, A. H. (2011). Regression for compositional data by using distributions defined on the hypersphere. *Journal of the Royal Statistical Society, Series B, Statistical Methodology*, 73:351–375.

Scealy, J. L. and Welsh, A. H. (2014). Fitting Kent models to compositional data with small concentration. *Statistics and Computing*, 24(2):165–179.

Song, P. X. and Tan, M. (2000). Marginal models for longitudinal continuous proportional data. *Biometrics*, 56(2):496–502.

Stephens, M. A. (1982). Use of the von Mises distribution to analyse continuous proportions. *Biometrika*, 69(1):197–203.

Tanabe, K. and Sagae, M. (1992). An exact Cholesky decomposition and the generalized inverse of the variance-covariance matrix of the multinomial distribution, with applications. *Journal of the Royal Statistical Society, Series B, Methodological*, 54:211–219.

Tsagris, M. (2014). Zero adjusted Dirichlet regression for compositional data with zero values present. *ArXiv e-prints*.

Tsagris, M. (2015). Regression analysis with compositional data containing zero values. *Chilean Journal of Statistics*, 6:45–57.

Tsagris, M. T., Preston, S., and Wood, A. T. A. (2011). A data-based power transformation for compositional data. In Egozcue, J., Tolosana-Delgado, R., and Ortego, M. I., editors, *Proceedings of CoDaWork'11: 4th International Workshop on Compositional Data Analysis*.

Warton, D. I. and Guttorp, P. (2011). Compositional analysis of overdispersed counts using generalized estimating equations. *Environmental and Ecological Statistics*, 18(3):427–446.

Wedderburn, R. (1974). Quasi-likelihood functions, generalized linear models and the Gauss-Newton method. *Biometrika*, 61:13–22.

Wedderburn, R. (1976). On the existence and uniqueness of the maximum likelihood estimates for certain generalized linear models. *Biometrika*, 63:27–32.

White, H. (1980). A heteroskedasticity-consistent covariance matrix estimator and a direct test for heteroskedasticity. *Econometrica*, 48(4):817–38.

White, H. (1982). Maximum likelihood estimation of misspecified models. *Econometrica*, 50(1):1–25.

Zadora, G., Neocleous, T., and Aitken, C. (2010). A two-level model for evidence evaluation in the presence of zeros. *Journal of Forensic Sciences*, 55(2):371–384.

Zeger, S. L. (1988). The analysis of discrete longitudinal data: Commentary. *Statistics in Medicine*, 7:161–168.

Zhang, B. (2013). *On Compositional Data Modeling and Its Biomedical Applications*. Dissertation, Columbia University.

Zhao, L. P., Prentice, R. L., and Self, S. G. (1992). Multivariate mean parameter estimation by using a partly exponential model. *Journal of the Royal Statistical Society, Series B, Methodological*, 54:805–811.