A Directional mixed effects model for compositional expenditure data

J. L. Scealy[1,*] and A. H. Welsh[2]

[1] *Research School of Finance, Actuarial Studies and Statistics, Australian National University, Canberra ACT 2601, Australia*

[2] *Mathematical Sciences Institute, Australian National University, Canberra ACT 2601, Australia*

[*] *Corresponding author's email: janice.scealy@anu.edu.au.*

Abstract

Compositional data are vectors of proportions defined on the unit simplex and this type of constrained data occur frequently in Government surveys. It is also possible for the compositional data to be correlated due to the clustering or grouping of the observations within small domains or areas. We propose a new class of mixed model for compositional data based on the Kent distribution for directional data, where the random effects also have Kent distributions. One useful property of the new directional mixed model is that the marginal mean direction has a closed form and is interpretable. The random effects enter the model in a multiplicative way via the product of a set of rotation matrices and the conditional mean direction is a random rotation of the marginal mean direction. In small area estimation settings the mean proportions are usually of primary interest and these are shown to be simple functions of the marginal

mean direction. For estimation we apply a quasi-likelihood method which results in solving a new set of generalised estimating equations and these are shown to have low bias in typical situations. For inference we use a nonparametric bootstrap method for clustered data which does not rely on estimates of the shape parameters (shape parameters are difficult to estimate in Kent models). We analyse data from the 2009-10 Australian Household Expenditure Survey CURF (confidentialised unit record file). We predict the proportions of total weekly expenditure on food and housing costs for households in a chosen set of domains. The new approach is shown to be more tractable than the traditional approach based on the logratio transformation.

Keywords: Clustered data; Generalised estimating equations; Repeated measurements; Small area estimation.

1 Introduction

Government surveys are typically designed to provide reliable estimates of population means or totals in large regions. Often it is also of interest to produce estimates in smaller domains or areas, but sample sizes may not be large enough to calculate reliable direct estimates. In this case it is usual to produce model based survey estimates which borrow strength from other data sources such as Census, administrative data or other auxiliary variables to reduce the mean squared errors (Rao, 2003; Pfeffermann, 2013). Model based estimates are commonly constructed by fitting mixed effects models including either linear or generalised linear mixed models and then calculating empirical best predictions for the domains of interest (e.g. Jiang and Lahiri, 2006; Jiang, 2007). The response variables in these models can sometimes be vectors constrained to sum to a known constant value (e.g. Militino et al, 2012; Molina et al, 2007).

In this paper we focus on modelling compositional household expenditure survey data from the 2009-10 Australian Household Expenditure Survey (HES) Confidentialised Unit Record File (CURF). The CURF is a data file released by the Australian Bureau of Statistics (ABS) with identifiers and some other data items removed or reduced in order to protect the confidentiality of respondents. The HES collects expenditure and demographic information by personal interview from private dwellings across Australia and dwellings are selected through a stratified, multistage cluster sampling design covering 97% of people living in Australia (very remote areas are excluded). Our analysis is based on the household level file which contains 9774 households in total. For further details regarding the CURF or HES see the user guide ABS (2012).

The response variable is the weekly household proportion of expenditure on housing (u_1), food (u_2) and other (u_3). Housing includes expenditure on current housing costs, domestic fuel and power, household furnishings and equipment, household services and operation, mortgage repayments and other capital housing costs. Food includes both food and non-alcoholic beverages and other represents the remaining expenditures. The responses $\boldsymbol{u} = (u_1, u_2, u_3)^T$ are compositions defined on the unit simplex $\boldsymbol{\Delta}^2$, where

$$\boldsymbol{\Delta}^{p-1} = \{(u_1, u_2, \ldots, u_p)^T : \quad u_k \geq 0 \quad (k = 1, 2, \ldots, p), \quad \sum_{k=1}^{p} u_k = 1\}$$

and the data $\boldsymbol{u}$ are also clustered in small domains. The explanatory variable is the logarithm of total weekly household expenditure (EXPTL). The left two plots in Figure 1 plot the observed housing and food proportions versus log (EXPTL) and the right two plots highlight two specific clusters. These plots show that the data are highly variable and there are possibly non-linear relationships.

Militino et al. (2012) fit a P-spline model to clustered compositional food expen-

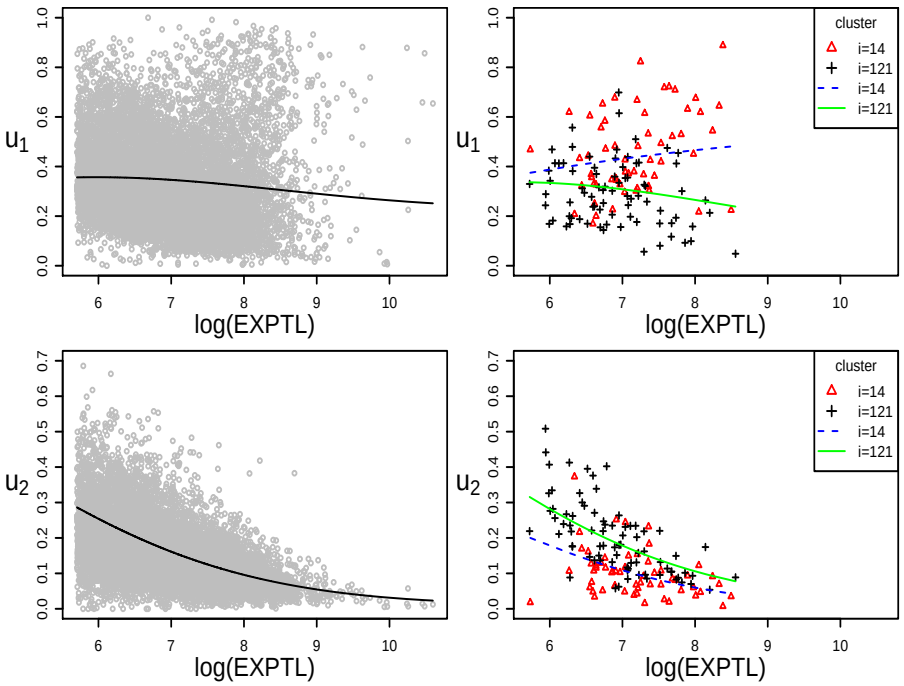

Figure 1: top left: $u_{1,ij}$ versus log (EXPTL) and estimated marginal mean curve; top right: $u_{1,ij}$ versus log (EXPTL) for clusters $i = 14$ and $i = 121$ including predicted conditional mean curves; bottom left: $u_{2,ij}$ versus log (EXPTL) and estimated marginal mean curve; bottom right: $u_{2,ij}$ versus log (EXPTL) for clusters $i = 14$ and $i = 121$ including predicted conditional mean curves.

diture data from the 2006 Spanish Household Budget Survey. The response variable in this model was the annual household percentage of expenditure on food and the explanatory variable was total annual expenditure. The clusters (small areas) of interest in the Spanish case were based on 52 provinces cross classified with different household size groups leading to 208 small areas. We use a similar set of variables in our models, but the available geography for defining our clusters is limited, so instead we use small domains based on the cross classification of three variables: state or territory of usual residence (STATEHBC) by area of usual residence (METHHC) by life cycle group (LIFECYCH) giving $m = 156$ small domains/clusters. The variable LIFECYCH is a fine level data item containing 12 categories based on family composition and the age distribution of household members. The cluster sizes n_i

range from 3 to 408 with median and mean 31 and 55 respectively, so the data are highly unbalanced.

The P-spline model of Militino et. al. (2012) assumed that the unit level residuals had a Gaussian distribution. One problem with this approach is that the simplex constraints are not taken into account. As seen in both Figure 1 and Militino et. al. (2012) Figure 2, p 2940, the proportions are distributed quite close to the zero boundary possibly leading to a non-negligible probability of negative values under a Gaussian model. One popular way of handling the sum to one constraint on compositional data is to apply a logratio transformation (Aitchison, 1982) and, to account for any clustering, to define a multivariate linear mixed model in the logratio space. However, as we will show, a disadvantage of this approach is that predictions of the proportions on the original scale are difficult to obtain under this model.

Various alternative methods have been proposed in the literature for treating compositional data; for a recent review see Scealy and Welsh (2014b). The idea of modelling compositional data using distributions for directional data has a long history (e.g. Stephens, 1982). If $\boldsymbol{y} = (y_1, y_2, \ldots, y_p)^T \in S^{p-1}$, where $S^{p-1} = \{\boldsymbol{y} \in \mathbb{R}^p : \parallel \boldsymbol{y} \parallel = 1\}$ denotes the unit hypersphere, then $\boldsymbol{u} = \boldsymbol{y}^2 = (y_1^2, y_2^2, \ldots, y_p^2)^T \in \boldsymbol{\Delta}^{p-1}$ is on the simplex. Scealy and Welsh (2011) proposed modelling $\boldsymbol{y}$ using a Kent distribution, which results in

$$| \boldsymbol{y} | = (| y_1 |, | y_2 |, \ldots, | y_p |)^T = (\sqrt{u_1}, \sqrt{u_2}, \ldots, \sqrt{u_p})^T = \sqrt{\boldsymbol{u}}$$

having a 'folded' Kent distribution and $\boldsymbol{u}$ having a 'squared' Kent distribution. The Kent distribution has a flexible covariance structure and is a special case of the Fisher-Bingham distribution (Kent, 1982; Mardia and Jupp, 2000 pp 176-177). In this paper we develop new mixed effects models for $\boldsymbol{y}$ based on Kent distributions and in estimation we use the approximation $\boldsymbol{y} = \sqrt{\boldsymbol{u}}$ which ignores the folding. This

is a valid approach as long as there is not a large amount of data distributed on or close to the zero boundaries with high variance (Scealy and Welsh, 2014a).

Scealy and Welsh (2011) proposed a regression model which models the mean direction of the Kent distribution as a function of covariates and they assumed the other parameters in the model were constant. In this paper we extend these regression models to account for clustered data. The idea is to let random effects define rotations of the mean direction so that, for example, the mean direction for a cluster is a random rotation of the overall mean direction. For most nonlinear mixed models and generalised linear mixed models, the marginal distribution does not exist in a closed form and the marginal moments involve integrals which are difficult to evaluate. Similar to the marginal mean for the Gaussian linear mixed model, we show that the marginal mean direction for the directional mixed model proposed in this paper has a closed form and we are able to estimate it consistently and reasonably efficiently using a quasi-likelihood moment based method (e.g. Heyde, 1997; McCullagh and Nelder, 1989, Chapter 10). Importantly the moment based estimation method does not require estimating the shape parameters which are difficult to estimate unbiasedly in Kent models when the variability is large (e.g. Scealy and Welsh, 2014a). We show that inferences for the moments can be obtained by direct application of the generalised cluster bootstrap method (Field et al., 2010).

There is not much literature on parametric models for correlated data on spheres or on more general manifolds. Two examples are Haddou et al. (2010) and Barry and Bowman (2008), but both of these effectively rely on large concentration (small variance) approximations. One advantage of our new directional model and our new moment based estimators is that they are valid even when the residual variance is not small, which is the case for the HES data. Similar to Militino et. al. (2012), our new directional mixed effects model is able to fit different non-linear predicted curves

in each cluster on the original scale (see Figure 1 right hand plots). One further advantage of our new model is that the marginal mean proportions have a simple parametric form and this aids interpretation. The new directional mixed model is also better able to account for the boundaries of the sample space. The majority of the observations on the square root scale are transformed away from the zero boundary (compare Figure 1 with Figure 2 in Section 6) and our predicted mean proportions are always defined on the simplex.

The rest of the paper is organised as follows. In Section 2 we define a new set of directional mixed effects models for compositional data. In Sections 3 and 4 we calculate the conditional and marginal moments and derive a useful asymptotic result. We discuss estimation and inference in Section 5 and we apply the new estimators to analyse the 2009-10 Australian HES CURF in Section 6. We also include a comparison of our new model with the logratio method and other approaches in Section 6. A simulation study is then undertaken in Section 7 to examine the properties of the estimators. We conclude with a brief discussion in Section 8.

2 Directional mixed effects models

The sample consists of compositional vector responses $\{\boldsymbol{u}_{ij} \in \boldsymbol{\Delta}^{p-1} : j = 1, 2, ..., n_i;\ i = 1, 2, ..., m\}$, where i denotes clusters, j represents units within clusters and n_i is the sample size within the ith cluster. There is also a set of covariates $\{\boldsymbol{x}_{ij} \in \mathbb{R}^q : j = 1, 2, ..., n_i;\ i = 1, 2, ..., m\}$ measured. Given $\boldsymbol{x}_{ij}$, the compositional data vectors $\boldsymbol{u}_{ij}$ are independent between different clusters, but possibly correlated within clusters. **To model the correlation within clusters we define a set of random effects at the cluster level.** Let $\boldsymbol{b}_i$ be a vector of random effects associated with cluster i; these are independent between clusters. The compositional data are each ~~linked with~~

mapped to a directional data vector via $\boldsymbol{u}_{ij} = \boldsymbol{y}_{ij}^2 = (y_{1,ij}^2, y_{2,ij}^2, \ldots, y_{p,ij}^2)^T$, where **the vector** $\boldsymbol{y}_{ij} = (y_{1,ij}, y_{2,ij}, \ldots, y_{p,ij})^T \in S^{p-1}$ ~~the unit hypersphere~~. We model the conditional distribution of $\boldsymbol{y}_{ij}$ given $\boldsymbol{b}_i$ and $\boldsymbol{x}_{ij}$ resulting in a new directional mixed effects model for compositional data. In this conditional model the observations $\boldsymbol{y}_{ij}$ are independent given $\boldsymbol{x}_{ij}$ and $\boldsymbol{b}_i$.

Specifically we assume that each $\boldsymbol{y}_{ij}|\boldsymbol{x}_{ij}, \boldsymbol{b}_i$ follows a Kent distribution on S^{p-1} with density

$$f(\boldsymbol{y}_{ij}|\boldsymbol{x}_{ij}, \boldsymbol{b}_i) \propto \exp\left\{\kappa \boldsymbol{\zeta}(\boldsymbol{x}_{ij}, \boldsymbol{b}_i)^T \boldsymbol{y}_{ij} + \boldsymbol{y}_{ij}^T \boldsymbol{\Gamma}(\boldsymbol{x}_{ij}, \boldsymbol{b}_i) \boldsymbol{D} \boldsymbol{\Gamma}(\boldsymbol{x}_{ij}, \boldsymbol{b}_i)^T \boldsymbol{y}_{ij}\right\}, \tag{1}$$

where $\kappa > 0$ and $\boldsymbol{\beta} = (\beta_2, \beta_3, \ldots, \beta_{p-1})^T \in \mathbb{R}^{p-2}$ are shape parameters satisfying

$$\frac{\kappa}{2} > \beta_2 \geq \beta_3 \geq \ldots \geq \beta_{p-1} \geq -\sum_{m=2}^{p-1} \beta_m, \tag{2}$$

and $\boldsymbol{D}$ is a diagonal matrix with $(0, \boldsymbol{\beta}^T, -\sum_{m=2}^{p-1} \beta_m)^T$ on the diagonal. The normalising constant in the above conditional density is a function only of the shape parameters. The $p \times p$ orthogonal matrix $\boldsymbol{\Gamma}(\boldsymbol{x}_{ij}, \boldsymbol{b}_i)$ contains the location parameters and the first column of $\boldsymbol{\Gamma}(\boldsymbol{x}_{ij}, \boldsymbol{b}_i)$ is $\boldsymbol{\zeta}(\boldsymbol{x}_{ij}, \boldsymbol{b}_i) \in S^{p-1}$, the conditional mean direction. This model assumes that the location parameters are functions of both $\boldsymbol{x}_{ij}$ and $\boldsymbol{b}_i$ and the shape parameters are constant. This is analogous to modelling the conditional mean as a function of both the covariates and random effects in linear mixed effects models and setting the variance components to constant values.

2.1 Random intercept model

~~In mixed effects models the random effects are usually assumed to have zero mean and similarly we assume that $\boldsymbol{b}_i = (b_{1,i}, b_{2,i}, \ldots, b_{p,i})^T \in S^{p-1}$ has a standardised Kent~~

~~distribution with density~~

$$f\left(\boldsymbol{b}_i\right) \propto \exp\left\{\kappa_b b_{1,i} + \boldsymbol{b}_{L,i}^T \boldsymbol{K}_b^* \text{diag}\left(\boldsymbol{\beta}_b^T, -\sum_{m=2}^{p-1} \beta_{m,b}\right) \boldsymbol{K}_b^{*T} \boldsymbol{b}_{L,i}\right\},$$

~~where $\boldsymbol{b}_{L,i} = (b_{2,i}, b_{3,i}, \ldots, b_{p,i})^T$, $\boldsymbol{K}_b^*$ is a general $(p-1) \times (p-1)$ orthogonal parameter matrix and diag($\boldsymbol{w}$) denotes a diagonal matrix with the vector $\boldsymbol{w}$ on the diagonal. Here κ_b is the concentration for the random effects and $\boldsymbol{\beta}_b = (\beta_{2,b}, \beta_{3,b}, \ldots, \beta_{p-1,b})^T$ are the remaining $p-2$ shape parameters satisfying the condition analogous to (2). In this model for the random effects both the shape parameters and location parameters are constant. The mean direction of the random effects distribution is at the north pole i.e. $(1, 0, 0, \ldots, 0)^T$ and $\boldsymbol{K}_b^*$ contains the remaining location parameters.~~

The location matrix $\boldsymbol{\Gamma}(\boldsymbol{x}_{ij}, \boldsymbol{b}_i)$ associated with the conditional model is defined in a multiplicative way as the product of a set of orthogonal matrices of the same dimension. Let

$$\boldsymbol{\Gamma}(\boldsymbol{x}_{ij}, \boldsymbol{b}_i) = \boldsymbol{H}(\boldsymbol{x}_{ij}) \boldsymbol{R}(\boldsymbol{b}_i) \begin{pmatrix} 1 & \boldsymbol{0}^T \\ \boldsymbol{0} & \boldsymbol{K}^* \end{pmatrix}, \tag{3}$$

where the matrix $\boldsymbol{K}^*$ is a general constant $(p-1) \times (p-1)$ orthogonal parameter matrix and $\boldsymbol{H}(\boldsymbol{x}_{ij})$ and $\boldsymbol{R}(\boldsymbol{b}_i)$ are defined below. Similar to Scealy and Welsh (2011) we link a parameter vector $\boldsymbol{\mu} \in S^{p-1}$ with $p-1$ linear functions of $\boldsymbol{x}_{ij}$ by using the square root of the additive logistic transformation (Aitchison, 1986, p 113). That is, let

$$\mu_k(\boldsymbol{x}_{ij}) = \begin{cases} \left(1 + \sum_{m=1}^{p-1} \exp\left(\boldsymbol{a}_m^T \boldsymbol{x}_{ij}\right)\right)^{-\frac{1}{2}} & k = 1 \\ \exp\left(\frac{\boldsymbol{a}_{k-1}^T \boldsymbol{x}_{ij}}{2}\right)\left(1 + \sum_{m=1}^{p-1} \exp\left(\boldsymbol{a}_m^T \boldsymbol{x}_{ij}\right)\right)^{-\frac{1}{2}} & k = 2, 3, ..., p, \end{cases}$$

where $\mu_k(\boldsymbol{x}_{ij})$ is the kth component of $\boldsymbol{\mu}(\boldsymbol{x}_{ij})$ and $\boldsymbol{a} = \left(\boldsymbol{a}_1^T, \boldsymbol{a}_2^T, \ldots, \boldsymbol{a}_{p-1}^T\right)^T \in \mathbb{R}^{q(p-1)}$

is a vector of regression coefficients. Then define

$$\boldsymbol{H}(\boldsymbol{x}_{ij}) = \begin{pmatrix} \mu_1(\boldsymbol{x}_{ij}) & \boldsymbol{\mu}_L(\boldsymbol{x}_{ij})^T \\ \boldsymbol{\mu}_L(\boldsymbol{x}_{ij}) & \frac{1}{1+\mu_1(\boldsymbol{x}_{ij})}\boldsymbol{\mu}_L(\boldsymbol{x}_{ij})\boldsymbol{\mu}_L(\boldsymbol{x}_{ij})^T - \boldsymbol{I}_{p-1} \end{pmatrix},$$

where $\boldsymbol{\mu}_L(\boldsymbol{x}_{ij}) = (\mu_2(\boldsymbol{x}_{ij}), \mu_3(\boldsymbol{x}_{ij}), ..., \mu_p(\boldsymbol{x}_{ij}))^T$ and $\boldsymbol{I}_{p-1}$ is the $(p-1)\times(p-1)$ identity matrix, and define the $p \times p$ rotation matrix

$$\boldsymbol{R}(\boldsymbol{b}_i) = \begin{pmatrix} b_{1,i} & -\boldsymbol{b}_{L,i}^T \\ \boldsymbol{b}_{L,i} & \boldsymbol{I}_{p-1} - \frac{1}{1+b_{1,i}}\boldsymbol{b}_{L,i}\boldsymbol{b}_{L,i}^T \end{pmatrix}.$$

We assume that $\boldsymbol{b}_i = (b_{1,i}, b_{2,i}, \ldots, b_{p,i})^T \in S^{p-1}$ has a standardised Kent distribution with density

$$f(\boldsymbol{b}_i) \propto \exp\left\{\kappa_b b_{1,i} + \boldsymbol{b}_{L,i}^T \boldsymbol{K}_b^* \text{diag}\left(\boldsymbol{\beta}_b^T, -\sum_{m=2}^{p-1}\beta_{m,b}\right)\boldsymbol{K}_b^{*T}\boldsymbol{b}_{L,i}\right\}, \tag{4}$$

where $\boldsymbol{b}_{L,i} = (b_{2,i}, b_{3,i}, \ldots, b_{p,i})^T$, $\boldsymbol{K}_b^*$ is a general $(p-1)\times(p-1)$ orthogonal parameter matrix and diag($\boldsymbol{w}$) denotes a diagonal matrix with the vector $\boldsymbol{w}$ on the diagonal. Here κ_b is the concentration for the random effects and $\boldsymbol{\beta}_b = (\beta_{2,b}, \beta_{3,b}, \ldots, \beta_{p-1,b})^T$ are the remaining $p-2$ shape parameters satisfying the condition analogous to (2). In this model for the random effects both the shape parameters and location parameters are constant.

The mean direction of the random effects distribution is at the north pole i.e. $(1, 0, 0, \ldots, 0)^T$ **and $\boldsymbol{K}_b^*$ contains the remaining location parameters. This effectively centers the model at direction $\boldsymbol{\mu}(\boldsymbol{x}_{ij})$, since the first column in $\boldsymbol{\Gamma}(\boldsymbol{x}_{ij}, \boldsymbol{b}_i)$ in (3) will be proportional to $\boldsymbol{\mu}(\boldsymbol{x}_{ij})$ on average. This is analogous to assuming that the random effects have mean zero in linear mixed effects models.** From (3) it follows that the **individual** effect of **each $\boldsymbol{b}_i$** is

to rotate the location of the mean direction $\boldsymbol{\mu}(\boldsymbol{x}_{ij})$ on the surface of the hypersphere so that units in the same cluster with the same covariates will tend to have more similar locations than units in different clusters with the same covariates. Alternative choices of $\boldsymbol{R}(.)$ are also possible including reflection matrices or other types of orthogonal matrices, but the chosen rotation matrix leads to a model with nice asymptotic properties and is ~~easier~~ **easy** to interpret (see Section 4). If we omit the random effects and set $\boldsymbol{R}(\boldsymbol{b}_i) = \boldsymbol{I}_p$, the model reduces to the additive regression model of Scealy and Welsh (2011) **with the same conditional mean direction** $\boldsymbol{\mu}(\boldsymbol{x}_{ij})$.

An alternative valid model could also be obtained by omitting the matrix $\boldsymbol{R}(.)$, incorporating Gaussian random effects into the definition of $\boldsymbol{\mu}(\boldsymbol{x}_{ij})$ and having the random effects on the same scale as the fixed effect regression coefficients. Under this model the matrix $\boldsymbol{H}$ would be a function of both the fixed and random effects (a similar structure is used in the multinomial logit mixed model). The problem with this approach is that the marginal moments under this model are difficult to calculate and do not exist in closed form. The advantage of our model is that it has a convenient multiplicative structure, where the random and fixed effects in (3) are separate terms, simplifying the calculation of the marginal moments (see Section 3).

2.2 Random slope model

It is also possible to incorporate a random slope term into the directional mixed effects model, allowing more flexibility when modelling the between cluster variation. Let $\{z_{ij} \in \mathbb{R} : j = 1, 2, ..., n_i;\ i = 1, 2, ..., m\}$ be a set of standardised random variables, each based on a subset or function of $\boldsymbol{x}_{ij}$ for a given unit. The model defined in Section 2.1 is extended by increasing the dimension of $\boldsymbol{b}_i$ to include $p-1$ extra components, that is let $\boldsymbol{b}_i \in S^{2p-2}$ have a standardised Kent distribution **(mean direction at the north pole)** with $2p-2$ shape parameters and with orthogonal location matrix $\boldsymbol{K}_b^*$

of size $(2p-2)\times(2p-2)$. Let $\boldsymbol{\xi}_{ij} = (\xi_{1,ij}, \xi_{2,ij}, \ldots, \xi_{p,ij})^T$ be a vector associated with a given unit which is defined as $\xi_{1,ij} = b_{1,i}$ and $\xi_{r,ij} = b_{r,i} + z_{ij} b_{p+r-1,i}$ for $r = 2, 3, \ldots, p$. We then project $\boldsymbol{\xi}_{ij}$ onto the hypersphere using $\boldsymbol{\xi}^*_{ij} = (\xi^*_{1,ij}, \xi^*_{2,ij}, \ldots, \xi^*_{p,ij})^T = \boldsymbol{\xi}_{ij} / \parallel \boldsymbol{\xi}_{ij} \parallel$. The location matrix $\boldsymbol{\Gamma}(\boldsymbol{x}_{ij}, \boldsymbol{b}_i)$ in the conditional model for $\boldsymbol{y}_{ij}|\boldsymbol{x}_{ij}, \boldsymbol{b}_i$ is replaced with

$$\boldsymbol{\Gamma}(\boldsymbol{x}_{ij}, \boldsymbol{b}_i) = \boldsymbol{H}(\boldsymbol{x}_{ij}) \boldsymbol{R}(\boldsymbol{\xi}^*_{ij}) \begin{pmatrix} 1 & \mathbf{0}^T \\ \mathbf{0} & \boldsymbol{K}^* \end{pmatrix}, \tag{5}$$

where the rotation matrix $\boldsymbol{R}(.)$ has the same form as in Section 2.1 and the other orthogonal matrices remain unchanged. **Similar to the random intercept model, it follows that the mean direction of $\boldsymbol{\xi}^*_{ij}$ given z_{ij} is at the north pole and this model is also centered at $\boldsymbol{\mu}(\boldsymbol{x}_{ij})$ on average.** This is the simplest random slope model. It is also possible to extend this model to include multiple random slopes by allowing z_{ij} to be a vector and by increasing the dimension of $\boldsymbol{b}_i$ accordingly.

3 Moments

A valuable property of the new directional mixed effects models proposed in Sections 2.1 and 2.2 is that the marginal moments exist in both an interpretable and a convenient form. The results in the next two sections are based on the model defined in Section 2.2 (expressions for the simpler model in Section 2.1 are omitted but can easily be derived in a similar way). Define

$$\boldsymbol{y}^*_{ij} = (y^*_{1,ij}, y^*_{2,ij}, \ldots, y^*_{p,ij})^T = \boldsymbol{R}(\boldsymbol{\xi}^*_{ij})^T \boldsymbol{H}(\boldsymbol{x}_{ij})^T \boldsymbol{y}_{ij},$$

for $j = 1, 2, \ldots, n_i$ and $i = 1, 2, \ldots, m$. These transformed observations are effectively a set of unit level residuals and each follow standardised Kent distributions with shape parameters κ and $\boldsymbol{\beta}$, mean direction at the north pole and the remaining location

parameters form the $(p-1)\times(p-1)$ orthogonal matrix $\boldsymbol{K}^*$. From Kent distribution properties, $\mathrm{E}\,(y^*_{1,ij}) = \psi(\kappa,\beta) = \psi$, $\psi > 0$ and applying the moment results in Scealy and Welsh (2011),

$$\mathrm{E}\,(\boldsymbol{y}_{ij}|\boldsymbol{x}_{ij},\boldsymbol{b}_i) = \psi\boldsymbol{H}(\boldsymbol{x}_{ij})\boldsymbol{\xi}^*_{ij}, \quad j = 1,2,\ldots,n_i, \quad i = 1,2,\ldots,m.$$

From symmetry properties of the Kent distribution (Scealy, 2010), it follows that conditional on z_{ij}, both $\mathrm{E}(\boldsymbol{\xi}_{L,ij}) = \boldsymbol{0}$ and $\mathrm{E}(\boldsymbol{\xi}_{L,ij}/\parallel \boldsymbol{\xi}_{ij} \parallel) = \boldsymbol{0}$, where $\boldsymbol{\xi}_{L,ij} = (\xi_{2,ij},\xi_{3,ij},\ldots,\xi_{p,ij})^T$ and hence

$$\mathrm{E}\,(\boldsymbol{y}_{ij}|\boldsymbol{x}_{ij}) = \psi\,\mathrm{E}\,(\xi^*_{1,ij}|z_{ij})\boldsymbol{\mu}(\boldsymbol{x}_{ij}), \quad j = 1,2,...,n_i, \quad i = 1,2,...,m. \tag{6}$$

Therefore the conditional mean direction is $\boldsymbol{\zeta}(\boldsymbol{x}_{ij},\boldsymbol{b}_i) = \boldsymbol{H}(\boldsymbol{x}_{ij})\boldsymbol{\xi}^*_{ij}$ and the marginal mean direction is $\boldsymbol{\mu}(\boldsymbol{x}_{ij})$ and importantly the marginal mean direction exists in a closed form. The marginal mean direction coincides with the mean direction in the regression model of Scealy and Welsh (2011).

Again applying the moment results in Scealy and Welsh (2011), the conditional second order moments are

$$\mathrm{E}\,(\boldsymbol{y}_{ij}\boldsymbol{y}^T_{ij}|\boldsymbol{x}_{ij},\boldsymbol{b}_i) = \boldsymbol{H}(\boldsymbol{x}_{ij})\boldsymbol{R}(\boldsymbol{\xi}^*_{ij})\boldsymbol{K}\boldsymbol{D}^*\boldsymbol{K}^T\boldsymbol{R}(\boldsymbol{\xi}^*_{ij})^T\boldsymbol{H}(\boldsymbol{x}_{ij})^T, \tag{7}$$

where $\boldsymbol{D}^*$ is a diagonal matrix which is a function of κ and $\boldsymbol{\beta}$ only and $\boldsymbol{K} = \text{block diag}(1,\boldsymbol{K}^*)$ is a $p\times p$ block diagonal matrix. From symmetry properties of the Kent distribution, including $\mathrm{E}\,(\xi_{1,ij}\boldsymbol{\xi}_{L,ij}/\parallel \boldsymbol{\xi}_{ij} \parallel^2) = \boldsymbol{0}$ (Scealy, 2010), we obtain

$$\mathrm{E}\,(\boldsymbol{y}_{ij}\boldsymbol{y}^T_{ij}|\boldsymbol{x}_{ij}) = \boldsymbol{H}(\boldsymbol{x}_{ij})\text{block diag}\left(1-\text{trace}\left(\boldsymbol{V}^{(i)}_{jj}\right),\boldsymbol{V}^{(i)}_{jj}\right)\boldsymbol{H}(\boldsymbol{x}_{ij})^T, \tag{8}$$

for $j = 1,2,\ldots,n_i$ and $i = 1,2,\ldots,m$, where $\boldsymbol{V}^{(i)}_{jj}$ is a $(p-1)\times(p-1)$ matrix

which is a function of $\boldsymbol{K}^*$, $\boldsymbol{K}_b^*$, κ, $\boldsymbol{\beta}$, κ_b, $\boldsymbol{\beta}_b$ and z_{ij}. The marginal moments (8) are particularly useful in a compositional data context because the diagonal is $\mathrm{E}(\boldsymbol{u}_{ij}|\boldsymbol{x}_{ij})$, the expected marginal proportions.

Let $\boldsymbol{e}_{ij} = (e_{1,ij}, e_{2,ij}, \ldots, e_{p,ij})^T$ represent the total residual error term for a given unit including the random effect components. That is,

$$\boldsymbol{e}_{ij} = \boldsymbol{H}(\boldsymbol{x}_{ij})^T \boldsymbol{y}_{ij} = \boldsymbol{R}(\boldsymbol{\xi}_{ij}^*)\boldsymbol{y}_{ij}^*.$$

From (6) and the fact that $\boldsymbol{H}(\boldsymbol{x}_{ij})$ is an orthogonal matrix it follows that $\mathrm{E}(\boldsymbol{e}_{ij}|z_{ij}) = \psi\,\mathrm{E}\,(\xi_{1,ij}^*|z_{ij})(1,0,0,\ldots,0)^T$ showing that the lower $p-1$ components each have zero mean. We now explore the second order moment structure of the residual error terms $\boldsymbol{e}_{L,ij} = (e_{2,ij}, e_{3,ij}, \ldots, e_{p,ij})^T$ (omitting the first component) within and between clusters. The data in different clusters are independent given the covariates so $\mathrm{E}(\boldsymbol{e}_{L,ij}\boldsymbol{e}_{L,rk}^T|z_{ij}, z_{rk}) = \boldsymbol{0}$ for $i \neq r$ and from (8) $\mathrm{E}(\boldsymbol{e}_{L,ij}\boldsymbol{e}_{L,ij}^T|z_{ij}) = \boldsymbol{V}_{jj}^{(i)}$. We assume all $\boldsymbol{y}_{ij}^*$ terms and random effects are independent of one another, which implies for $j \neq k$,

$$\mathrm{E}(\boldsymbol{e}_{L,ij}\boldsymbol{e}_{L,ik}^T|z_{ij}, z_{ik}) = \psi^2\,\mathrm{E}(\boldsymbol{\xi}_{L,ij}^*\boldsymbol{\xi}_{L,ik}^{*T}|z_{ij}, z_{ik}) = \boldsymbol{V}_{jk}^{(i)},$$

where $\boldsymbol{\xi}_{L,ij}^* = (\xi_{2,ij}^*, \xi_{3,ij}^*, \ldots, \xi_{p,ij}^*)^T$. The $(p-1)\times(p-1)$ matrix $\boldsymbol{V}_{jk}^{(i)}$ is a function of ψ, κ_b, $\boldsymbol{\beta}_b$, $\boldsymbol{K}_b^*$, z_{ij} and z_{ik}.

4 Asymptotic results

We now derive some useful asymptotic results for the mixed model defined in Section 2.2. Assume that the following limit conditions hold for the shape parameters κ and $\boldsymbol{\beta}$

$$\kappa \to \infty, \quad \frac{\beta_m}{\kappa} \to d_m \quad \text{and} \quad \frac{1}{2} > d_2 \geq d_3 \ldots \geq d_{p-1} \geq -\left(\sum_{m=2}^{p-1} d_m\right) \tag{9}$$

and analogous conditions also hold for κ_b and $\boldsymbol{\beta}_b$. Then applying Scealy and Welsh (2011) Theorem 3, we obtain $\boldsymbol{y}^*_{ij} = (1, \boldsymbol{y}^{*T}_{L,ij})^T + \boldsymbol{O}_p(\kappa^{-1})$ and $\boldsymbol{b}_i = (1, \boldsymbol{b}^T_{L,i})^T + \boldsymbol{O}_p(\kappa_b^{-1})$, where $\boldsymbol{y}^*_{L,ij} = (y^*_{2,ij}, y^*_{3,ij}, \ldots, y^*_{p,ij})^T$ and $\boldsymbol{b}_{L,i} = (b_{2,i}, b_{3,i}, \ldots, b_{2p-1,i})^T$. The following approximations also hold:

$$\boldsymbol{y}^*_{L,ij} \sim N(\boldsymbol{0}, \boldsymbol{\Sigma}) \quad \text{and} \quad \boldsymbol{b}_{L,i} \sim N(\boldsymbol{0}, \boldsymbol{\Sigma}_b), \tag{10}$$

where $\boldsymbol{\Sigma}$ is a known function of κ, $\boldsymbol{\beta}$ and $\boldsymbol{K}^*$ and $\boldsymbol{\Sigma}_b$ is a known function of κ_b, $\boldsymbol{\beta}_b$ and $\boldsymbol{K}^*_b$ (Scealy and Welsh, 2011). The following theorem provides a useful asymptotic result for the directional mixed effects model.

Theorem 4.1. *Assume that $\boldsymbol{y}_{ij}$ follows a directional mixed effects model as defined in Section 2.2 and assume that the limit conditions (9) hold for κ and $\boldsymbol{\beta}$ and analogous conditions also hold for κ_b and $\boldsymbol{\beta}_b$. Additionally, assume that there exists a constant $0 < c^* < \infty$ such that $\lim_{\kappa\to\infty} \frac{\kappa_b}{\kappa} = c^*$ and $z_{ij} = O_p(1)$. Then $\kappa^{\frac{1}{2}} \boldsymbol{e}_{L,ij} - \kappa^{\frac{1}{2}} \left(\boldsymbol{\xi}_{L,ij} + \boldsymbol{y}^*_{L,ij}\right) = \boldsymbol{O}_p(\kappa^{-1})$.*

Proof. Under the limit conditions $\xi^*_{1,ij} = b_{1,i} + O_p(\kappa_b^{-1})$ and $\boldsymbol{\xi}^*_{L,ij} = \boldsymbol{\xi}_{L,ij} + \boldsymbol{O}_p(\kappa_b^{-\frac{3}{2}})$. Then substituting these expressions into the definition of $\boldsymbol{R}(\boldsymbol{\xi}^*_{ij})$ gives

$$\boldsymbol{e}_{ij} = \begin{pmatrix} 1 + O_p(\kappa_b^{-1}) & -\boldsymbol{\xi}^T_{L,ij} + \boldsymbol{O}_p(\kappa_b^{-\frac{3}{2}}) \\ \boldsymbol{\xi}_{L,ij} + \boldsymbol{O}_p(\kappa_b^{-\frac{3}{2}}) & \boldsymbol{I}_{p-1} + \boldsymbol{O}_p(\kappa_b^{-1}) \end{pmatrix} \begin{pmatrix} 1 + O_p(\kappa^{-1}) \\ \boldsymbol{y}^*_{L,ij} \end{pmatrix}$$

The limit conditions imply that κ_b and κ approach ∞ together at the same rate, which implies $O_p(\kappa^{-q}) = O_p(\kappa_b^{-q})$ for any $q > 0$ and hence $\boldsymbol{e}_{L,ij} = \boldsymbol{\xi}_{L,ij} + \boldsymbol{y}^*_{L,ij} + \boldsymbol{O}_p(\kappa^{-\frac{3}{2}})$. □

A consequence of (10) and Theorem 4.1 is that the residual $\boldsymbol{e}_{L,ij}$ has a marginal Gaussian distribution asymptotically.

5 Estimation and Inference

5.1 Estimation

The marginal distribution for the Kent mixed model involves multidimensional integrals which are difficult to evaluate and so directly maximising the marginal loglikelihood is not straightforward. Similar problems occur in other directional data mixed model settings. For example Haddou et al. (2010) defined a non-linear mixed effects model for 3×3 rotation matrix data and applied large κ approximations coupled with a further integral approximation to calculate the marginal loglikelihood. In many applications prediction of $\mathrm{E}\left(\boldsymbol{u}_{ij}|\boldsymbol{x}_{ij}\right)$ is of primary interest and this is a function of the parameters $\boldsymbol{a}$ and $\boldsymbol{V}_{jj}^{(i)}$, so we focus on the estimation of $\boldsymbol{a}$ and $\boldsymbol{V}_{jj}^{(i)}$ for $j = 1, 2, \ldots, n_i$ and $i = 1, 2, \ldots, m$. Instead of approximate maximum likelihood estimation, we propose a quasi-likelihood approach for estimating $\boldsymbol{a}$ since the marginal moments are available in a convenient form.

Scealy and Welsh (2011) proposed a preliminary estimate $\hat{\boldsymbol{a}}$ of $\boldsymbol{a}$ that is defined by the estimating equation

$$\sum_{i=1}^{m} \sum_{j=1}^{n_i} \frac{\partial \boldsymbol{\mu}(\boldsymbol{x}_{ij})}{\partial \boldsymbol{a}} \boldsymbol{y}_{ij} = \boldsymbol{0}. \tag{11}$$

This is also an unbiased estimating equation for the directional mixed effects model since $\boldsymbol{\mu}(\boldsymbol{x}_{ij})$ is the marginal mean direction. However, the estimator (11) may not be fully efficient for the mixed model because it ignores both the clustered nature of the data and the dependence within each vector $\boldsymbol{y}_{ij}$. Applying Scealy and Welsh (2011)

Theorem 5, $\frac{\partial \boldsymbol{\mu}(\boldsymbol{x}_{ij})}{\partial \boldsymbol{a}} \boldsymbol{\mu}(\boldsymbol{x}_{ij}) = \mathbf{0}$ and since $\boldsymbol{H}(\boldsymbol{x}_{ij})$ is orthogonal,

$$\begin{aligned} \frac{\partial \boldsymbol{\mu}(\boldsymbol{x}_{ij})}{\partial \boldsymbol{a}} \boldsymbol{y}_{ij} &= \frac{\partial \boldsymbol{\mu}(\boldsymbol{x}_{ij})}{\partial \boldsymbol{a}} \left(\boldsymbol{I}_p - \boldsymbol{\mu}(\boldsymbol{x}_{ij}) \boldsymbol{\mu}(\boldsymbol{x}_{ij})^T \right) \boldsymbol{y}_{ij} \\ &= \frac{\partial \boldsymbol{\mu}(\boldsymbol{x}_{ij})}{\partial \boldsymbol{a}} \boldsymbol{H}^*(\boldsymbol{x}_{ij}) \boldsymbol{H}^*(\boldsymbol{x}_{ij})^T \boldsymbol{y}_{ij}, \end{aligned} \tag{12}$$

where we define $\boldsymbol{H}(\boldsymbol{x}_{ij}) = (\boldsymbol{\mu}(\boldsymbol{x}_{ij}), \boldsymbol{H}^*(\boldsymbol{x}_{ij}))$ and $\boldsymbol{H}^*(\boldsymbol{x}_{ij})$ is the $p \times (p-1)$ submatrix of $\boldsymbol{H}(\boldsymbol{x}_{ij})$ with columns orthogonal to $\boldsymbol{\mu}(\boldsymbol{x}_{ij})$. As we now show, this new factorisation (12) based on the residuals $\boldsymbol{e}_{L,ij} = \boldsymbol{H}^*(\boldsymbol{x}_{ij})^T \boldsymbol{y}_{ij}$ is helpful when developing a more efficient set of estimators for $\boldsymbol{a}$.

Let $\boldsymbol{y}_i = (\boldsymbol{y}_{i1}^T, \ldots, \boldsymbol{y}_{in_i}^T)^T$, $\boldsymbol{\mu}_i = (\boldsymbol{\mu}(\boldsymbol{x}_{i1})^T, \ldots, \boldsymbol{\mu}(\boldsymbol{x}_{in_i})^T)^T$, $\boldsymbol{x}_i = (\boldsymbol{x}_{i1}^T, \ldots, \boldsymbol{x}_{in_i}^T)^T$ and $\boldsymbol{H}_i^* = \text{block diag}\,(\boldsymbol{H}^*(\boldsymbol{x}_{i1}), ..., \boldsymbol{H}^*(\boldsymbol{x}_{in_i}))$ for $i = 1, 2, ..., m$. Then define

$$\boldsymbol{V}_i = \mathrm{E}\left(\boldsymbol{H}_i^{*T} \boldsymbol{y}_i \boldsymbol{y}_i^T \boldsymbol{H}_i^* \right) = \begin{pmatrix} \boldsymbol{V}_{11}^{(i)} & \cdots & \boldsymbol{V}_{1n_i}^{(i)} \\ \vdots & \vdots & \vdots \\ \boldsymbol{V}_{n_i1}^{(i)} & \cdots & \boldsymbol{V}_{n_in_i}^{(i)} \end{pmatrix},$$

$\boldsymbol{H}^* = \text{block diag}\,(\boldsymbol{H}_1^*, ..., \boldsymbol{H}_m^*)$, $\boldsymbol{y} = \left(\boldsymbol{y}_1^T, ..., \boldsymbol{y}_m^T\right)^T$ and $\boldsymbol{x} = \left(\boldsymbol{x}_1^T, ..., \boldsymbol{x}_m^T\right)^T$. Now consider the class of estimating functions $\mathscr{G} = \left\{ \boldsymbol{G} = \boldsymbol{A}\boldsymbol{H}^{*T}\boldsymbol{y} \right\}$, where $\boldsymbol{A} = \boldsymbol{A}(\boldsymbol{a}, \boldsymbol{x})$ is a $q(p-1) \times n(p-1)$ matrix and $n = \sum_{i=1}^m n_i$.

Theorem 5.1. *Assume that $\boldsymbol{y}$ follows the random intercept directional mixed effects model as described in Section 2.1. The optimal estimating function within $\mathscr{G}$ for this model is*

$$\boldsymbol{G}^* = \frac{\partial \boldsymbol{\mu}}{\partial \boldsymbol{a}} \boldsymbol{H}^* \boldsymbol{V}^{-1} \boldsymbol{H}^{*T} \boldsymbol{y},$$

where $\boldsymbol{V} = \mathrm{E}\left(\boldsymbol{H}^{*T} \boldsymbol{y} \boldsymbol{y}^T \boldsymbol{H}^* \right)$.

Proof. Follows directly from Theorem 2.1 of Heyde (1997). A more detailed proof is given in Appendix A. □

The data in different clusters are independent given the covariates, so the optimal estimating equations for $\boldsymbol{a}$ given $\boldsymbol{V}_1, \boldsymbol{V}_2, \ldots, \boldsymbol{V}_m$ for the random intercept model are

$$\sum_{i=1}^{m} \frac{\partial \boldsymbol{\mu}_i}{\partial \boldsymbol{a}} \boldsymbol{H}_i^* \boldsymbol{V}_i^{-1} \boldsymbol{H}_i^{*T} \boldsymbol{y}_i = \boldsymbol{0}. \tag{13}$$

Note that these estimating equations are only approximately optimal for the random slope model defined in Section 2.2 when κ_b is large (see Appendix A for further details).

To estimate each $\boldsymbol{V}_{jj}^{(i)}$ we use the large κ and κ_b approximations derived in Section 4. These estimates are shown to be highly accurate even when κ and κ_b are only moderately large (see Section 7). In comparison estimates of the shape parameters based on large κ approximations are often quite biased (Scealy and Welsh, 2014a). The Gaussian approximation (valid when κ and κ_b are large) is

$$\boldsymbol{V}_{jj}^{(i)} \approx \boldsymbol{\Sigma} + \boldsymbol{A}_{ij} \boldsymbol{\Sigma}_b \boldsymbol{A}_{ij}^T \quad \text{and} \quad \boldsymbol{V}_{jk}^{(i)} \approx \boldsymbol{A}_{ij} \boldsymbol{\Sigma}_b \boldsymbol{A}_{ik}^T \quad (j \neq k), \tag{14}$$

where $\boldsymbol{A}_{ij} = (\boldsymbol{I}_{p-1} \quad z_{ij} \boldsymbol{I}_{p-1})$ for the random slope model or $\boldsymbol{A}_{ij} = \boldsymbol{I}_{p-1}$ for the random intercept model. The matrix $\boldsymbol{\Sigma}$ is a general $(p-1)$-dimensional covariance matrix and $\boldsymbol{\Sigma}_b$ is either a $(p-1)$ or $(2p-2)$-dimensional covariance matrix depending on which model is used (random intercept or random slope). Given $\boldsymbol{a}$, to obtain estimates of $\boldsymbol{V}_1, \boldsymbol{V}_2, \ldots, \boldsymbol{V}_m$, we maximise the approximate Gaussian loglikelihood

$$-\frac{n(p-1)}{2} \log(2\pi) - \frac{1}{2} \sum_{i=1}^{m} \left(\boldsymbol{y}_i^T \boldsymbol{H}_i^* \boldsymbol{V}_i^{-1} \boldsymbol{H}_i^{*T} \boldsymbol{y}_i + \log |\, \boldsymbol{V}_i \,| \right) \tag{15}$$

with respect to $\boldsymbol{\Sigma}$ and $\boldsymbol{\Sigma}_b$, where $\boldsymbol{\Sigma}$ and $\boldsymbol{\Sigma}_b$ are defined in (14).

To compute joint estimates of all the parameters, we iterate between solving (13) and maximising (15) a fixed number of times or until convergence. This is

similar to the iterative estimating equation approach defined by Jiang et al. (2007), although they use a different set of estimating equations. One advantage of this approximate approach is that estimation is straightforward since well established variance component estimation algorithms already available for linear mixed effects models can be used (e.g. Pinheiro and Bates, 2000). In many applications κ and κ_b will be moderately large and we do not expect to lose much efficiency when using this approximation for $\boldsymbol{V}_i$ when updating $\boldsymbol{a}$.

5.2 Inference

Scealy and Welsh (2011) used a parametric bootstrap for inference, however this approach relies on estimates of the shape parameters. Instead we suggest applying the generalised cluster bootstrap (Field et. al., 2010) which is a special case of the generalised bootstrap for estimating equations defined by Chatterjee and Bose (2005). The generalised cluster bootstrap has been studied by various authors in the context of Gaussian linear mixed models and robust estimation and has been shown to work well for both the variance components and mean parameters even in cases where the data are highly unbalanced (Field et al., 2010; Samanta and Welsh, 2013). The method works by replacing the estimating equations with weighted versions, that is solving

$$\sum_{i=1}^{m} w_{mi} \frac{\partial \boldsymbol{\mu}_i}{\partial \boldsymbol{a}} \boldsymbol{H}_i^* \boldsymbol{V}_i^{-1} \boldsymbol{H}_i^{*T} \boldsymbol{y}_i = \boldsymbol{0} \tag{16}$$

for the regression coefficients and maximising

$$-\frac{1}{2} \sum_{i=1}^{m} w_{mi} \left(\boldsymbol{y}_i^T \boldsymbol{H}_i^* \boldsymbol{V}_i^{-1} \boldsymbol{H}_i^{*T} \boldsymbol{y}_i + \log | \, \boldsymbol{V}_i \, | \right), \tag{17}$$

for the variance components $\boldsymbol{\Sigma}$ and $\boldsymbol{\Sigma}_b$, where $\boldsymbol{w}_m = (w_{m1}, w_{m2}, \ldots, w_{mm})^T$ has a multinomial distribution with parameters m and $m^{-1}(1, 1, \ldots, 1)^T$ (other distribu-

tions are also possible). The $\boldsymbol{w}_m$ are randomly sampled B times, where B is the number of bootstrap samples.

5.3 Modelling heteroscedasticity

The models defined in Sections 2.1 and 2.2 assume that $\boldsymbol{K}$, κ and $\boldsymbol{\beta}$ are constant and not dependent on the covariates. To account for heteroscedasticity, we suggest using the Gaussian approximation developed in Section 4. Following Pinheiro and Bates (2000, p 205), let $\boldsymbol{\Sigma} = \boldsymbol{\Sigma}(\boldsymbol{x}_{ij})$ be decomposed into the products

$$\boldsymbol{\Sigma}(\boldsymbol{x}_{ij}) = \sigma^2 \tilde{\boldsymbol{D}}(\boldsymbol{x}_{ij}) \boldsymbol{C} \tilde{\boldsymbol{D}}(\boldsymbol{x}_{ij}), \quad j = 1, 2, \ldots, n_i, \quad i = 1, 2, \ldots, m,$$

where σ^2 is a scalar parameter, $\tilde{\boldsymbol{D}}(\boldsymbol{x}_{ij})$ is a $(p-1)\times(p-1)$ diagonal matrix containing positive elements on the diagonal and $\boldsymbol{C}$ is a $(p-1) \times (p-1)$ correlation matrix. A general heteroscedastic model is obtained by defining

$$\tilde{\boldsymbol{D}}(\boldsymbol{x}_{ij}) = \text{diag}\left(g_1(\boldsymbol{x}_{ij}, \boldsymbol{\delta}), g_2(\boldsymbol{x}_{ij}, \boldsymbol{\delta}), \ldots, g_{p-1}(\boldsymbol{x}_{ij}, \boldsymbol{\delta})\right),$$

where $g_l(.)$ for $l = 1, 2, \ldots, p-1$ are a set of variance functions and $\boldsymbol{\delta}$ is a vector of variance parameters. To estimate the variance parameters, a modified version of the iterative estimating equation algorithm can be applied where the approximate loglikelihood (15) is maximised and each $\boldsymbol{V}_i$ is redefined to incorporate $\boldsymbol{\delta}$.

6 Modelling expenditure proportions in small domains

Let $u_{1,ij}$ represent the housing proportion in the jth dwelling within the ith small domain, $u_{2,ij}$ represent the food proportion in the jth dwelling within the ith small domain and let $u_{3,ij} = 1 - u_{1,ij} - u_{2,ij}$. After examining the distributions of these variables we decided to delete a small number of observations prior to modelling. We removed all observations with a negative value of either $u_{1,ij}$ or $u_{3,ij}$, all observations with EXPTL < 300 and one outlier, leaving a total sample size of $n = \sum_{i=1}^{m} n_i = 8594$. The reason we deleted small values of EXPTL is that the components of $\boldsymbol{u}_{ij}$ in this region are extremely highly variable over the whole range [0,1] and any model would have difficulty predicting in this region. Note that our data are based on weekly expenditures and not annual values as in the Spanish case, which may help to explain the high volatility. We choose the log transformation of EXPTL as the actual covariate in our models because this helps to reduce the high right skewness of the EXPTL distribution.

6.1 Directional heteroscedastic random slope model

We transform $\boldsymbol{u}_{ij}$ to $\boldsymbol{y}_{ij} = (y_{1,ij}, y_{2,ij}, y_{3,ij})^T = (\sqrt{u_{1,ij}}, \sqrt{u_{2,ij}}, \sqrt{u_{3,ij}})^T$ and work on the square root scale. The left hand plots in Figure 2 plot $y_{1,ij}$ versus log (EXPTL) and $y_{2,ij}$ versus log (EXPTL) and the right hand plots are similar, but only show two clusters $i = 14$ and $i = 121$ (the other clusters are omitted). Cluster $i = 14$ represents lone person aged under 35 in Sydney and $i = 121$ represents couple only, reference person aged 55 to 64 in Adelaide. These plots and other plots highlighting different clusters (not shown here), suggest that there are relationships between $\boldsymbol{y}_{ij}$ and EXPTL and these relationships may differ between clusters/groups. For example,

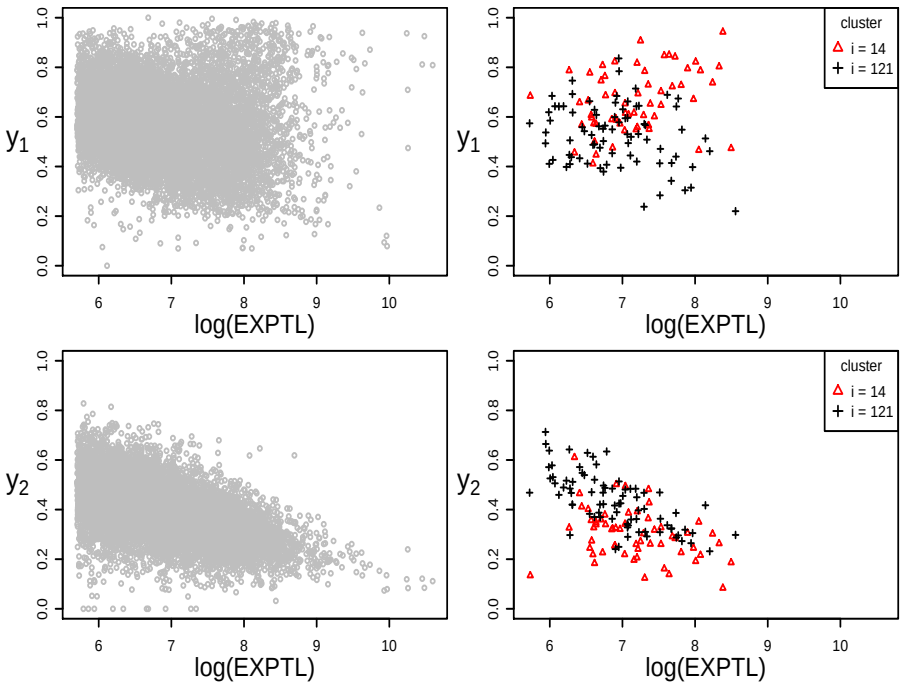

Figure 2: top left: $y_{1,ij}$ versus log (EXPTL); top right: $y_{1,ij}$ versus log (EXPTL) for clusters $i = 14$ and $i = 121$; bottom left: $y_{2,ij}$ versus log (EXPTL); bottom right: $y_{2,ij}$ versus log (EXPTL) for clusters $i = 14$ and $i = 121$.

$y_{2,ij}$ appears to on average decrease with EXPTL in both clusters $i = 14$ and $i = 121$, but at possibly different rates and cluster $i = 14$ observations generally have lower values of $y_{2,ij}$ than in cluster $i = 121$. Many clusters overlap in their distributions, so the between cluster variation is not expected to be large compared with the unit level residual variation. The component $y_{1,ij}$ is highly variable and its relationship with EXPTL is weak. The variation in $y_{2,ij}$ is moderately large when EXPTL is small and the variation is small when EXPTL is large, implying that a heteroscedastic model may be needed.

Motivated by the above discussion we fit the random slope model defined in Section 2.2 with heteroscedastic errors (see Section 5.3). The vector of covariates in the model is

$$\boldsymbol{x}_{ij} = (1, x_{ij})^T = \left(1, \frac{\log(\text{EXPTL}_{ij}) - m^*}{s^*}\right)^T,$$

where EXPTL_{ij} represents the ijth sample value of EXPTL and m^* and s^* are the sample mean and sample standard deviations respectively of the variable $\log(\text{EXPTL}_{ij})$ $i = 1, 2, \ldots, m$ and $j = 1, 2, \ldots, n_i$ (the covariates are standardised on the log scale). The regression coefficient parameter vectors are $\boldsymbol{a}_1 = (a_{11}, a_{12})^T$ and $\boldsymbol{a}_2 = (a_{21}, a_{22})^T$. We also set $z_{ij} = x_{ij}$ and use the same covariate in both the random and fixed parts of the model.

We apply the Gaussian approximation (14) to estimate the variance components and we use the following parametrisations

$$\boldsymbol{\Sigma}(\boldsymbol{x}_{ij}) = \sigma_1^2 \begin{pmatrix} v_{ij}^{\delta_1} & 0 \\ 0 & \sigma_2 v_{ij}^{\delta_2} \end{pmatrix} \begin{pmatrix} 1 & c_1 \\ c_1 & 1 \end{pmatrix} \begin{pmatrix} v_{ij}^{\delta_1} & 0 \\ 0 & \sigma_2 v_{ij}^{\delta_2} \end{pmatrix},$$

where $v_{ij} = c_2^{-1} \log(\text{EXPTL}_{ij})$ and $c_2 = 10.59663$ (this is the maximum value of $\log(\text{EXPTL}_{ij})$ in the sample) and $\boldsymbol{\Sigma}_b$ is a 4×4 symmetric matrix with r, sth element denoted by τ_{rs}. The parameter estimates are given in Table 1. The estimated standard errors and 95% confidence intervals were obtained from the nonparametric bootstrap method described in Section 5.2 with $B = 1000$ replications. To maximise (17) we used the lme function with standard settings from the R library nlme because it incorporates heteroscedasticity and correlation in the residuals. In 5% of cases the lme function reported a singular convergence error code and we discarded these cases before calculating the Monte Carlo estimates. The reason for the nonconvergence issue is the fact that the variance components are small and close to the zero boundary. The regression coefficients, σ_1, c_1, δ_1 and δ_2 are significantly different from zero giving evidence that the heteroscedastic model is needed. The variance components τ_{11}, τ_{22}, τ_{33} and τ_{44} are small, but are still significantly different from zero.

The left two plots in Figure 1 in the Introduction show the observed housing proportions $u_{1,ij}$ versus $\log(\text{EXPTL})$ and the observed food proportions $u_{2,ij}$ ver-

Table 1: Parameter estimates and bootstrap standard error and 95% confidence interval estimates.

	Estimate (SE)	Confidence Interval		Estimate (SE)	Confidence Interval
a_{11}	-0.748(0.032)	(-0.81,-0.68)	τ_{11}	0.00231(0.00033)	(0.0016,0.0029)
a_{12}	-0.324(0.024)	(-0.37,-0.27)	τ_{12}	0.000780(0.00022)	(0.00036,0.0012)
a_{21}	0.393(0.029)	(0.34,0.45)	τ_{13}	-0.000340(0.00010)	(-0.00054,-0.00013)
a_{22}	0.219(0.023)	(0.18,0.27)	τ_{14}	0.000748(0.00022)	(0.00031,0.0012)
σ_1	0.0495(0.0025)	(0.044,0.054)	τ_{22}	0.00194(0.00031)	(0.0013,0.0025)
σ_2	5.19(0.27)	(4.7,5.7)	τ_{23}	-0.000106(0.00010)	(-0.00032,0.000096)
δ_1	-1.73(0.12)	(-1.9,-1.5)	τ_{24}	0.000253(0.00022)	(-0.00021,0.00067)
δ_2	0.806(0.092)	(0.62, 0.98)	τ_{33}	0.000255(0.000044)	(0.00017,0.00034)
c_1	0.356(0.013)	(0.33,0.38)	τ_{34}	0.000297(0.00010)	(0.000088,0.00049)
			τ_{44}	0.00120(0.00029)	(0.00061,0.0017)

sus log (EXPTL). The curved lines correspond to the estimated expected marginal proportions given EXPTL. These estimates were obtained from the diagonal terms in (8), with the parameters replaced by their estimated values. The fit is good. In small area estimation problems it is of interest to produce predicted conditional means for each small area (e.g. Militino et al, 2012; Molina et al, 2007). The right hand plots in Figure 1 contain the observed values and the predicted conditional mean curves on the original scale for clusters $i = 14$ and $i = 121$. These predictions were obtained from the diagonal elements in (7), with the parameters replaced by their estimated values. Note that the matrix $\boldsymbol{K}\boldsymbol{D}^*\boldsymbol{K}^T$ for each $\boldsymbol{x}_{ij}$ was replaced by block diag $\left(1 - \text{trace}\left(\hat{\boldsymbol{\Sigma}}\left(\boldsymbol{x}_{ij}\right)\right), \hat{\boldsymbol{\Sigma}}\left(\boldsymbol{x}_{ij}\right)\right)$, where $\hat{\boldsymbol{\Sigma}}\left(\boldsymbol{x}_{ij}\right)$ is the estimate of $\boldsymbol{\Sigma}\left(\boldsymbol{x}_{ij}\right)$. The prediction of each random effect $\boldsymbol{\xi}_{ij}^*$ denoted by $\hat{\boldsymbol{\xi}}_{ij}^*$ was calculated using linear mixed model empirical best predictors (e.g. Jiang, 2007, pp 142-150). This is a valid approximation since $\boldsymbol{\Sigma}_b$ is small and the predictions fit well.

To further examine the covariance structure and model assumptions we calculated the predicted unit level residuals

$$\hat{\boldsymbol{y}}_{ij}^* = (\hat{y}_{1,ij}^*, \hat{y}_{2,ij}^*, \hat{y}_{3,ij}^*)^T = \boldsymbol{R}\left(\hat{\boldsymbol{\xi}}_{ij}^*\right)^T \hat{\boldsymbol{H}}(\boldsymbol{x}_{ij})^T \boldsymbol{y}_{ij}$$

and the standardised lower components

$$(\tilde{y}^*_{2,ij}, \tilde{y}^*_{3,ij})^T = \left(\hat{\boldsymbol{\Sigma}}(\boldsymbol{x}_{ij})\right)^{-1/2} (\hat{y}^*_{2,ij}, \hat{y}^*_{3,ij})^T,$$

where $\boldsymbol{M}^{1/2}$ denotes the square root of a matrix $\boldsymbol{M}$. We examined many different residual plots (not shown here) including plots of $\tilde{y}^*_{2,ij}$ and $\tilde{y}^*_{3,ij}$ versus log (EXPTL) within each cluster. The points were distributed fairly randomly with no obvious patterns. Initially we fitted a random intercept model (see Section 2.1) to the data with no heteroscedasticity, but the residual plots revealed a nonconstant variance pattern and relationships between the residuals and log (EXPTL) within each cluster. Introducing the random slope term and incorporating heteroscedasticity corrected these problems.

Plots (a) and (b) in Figure 3 contain Gaussian quantile-quantile plots of $\tilde{y}^*_{2,ij}$ and $\tilde{y}^*_{3,ij}$ respectively. The estimated quantiles do deviate mildly from Gaussian, but this is not surprising since the sample size is large. It is unlikely that the Kent distribution assumption for $\boldsymbol{y}^*_{ij}$ will hold exactly, but importantly the estimators and the non parametric bootstrap method are still valid as long as the underlying moment assumptions hold and the residuals $y^*_{2,ij}$ and $y^*_{3,ij}$ have symmetric distributions. The distribution of $\tilde{y}^*_{2,ij}$ is very close to symmetric and $\tilde{y}^*_{3,ij}$ is mildly asymmetric, implying that the symmetry assumption is not overly violated. Further, because the distributions of the unit level residuals $y^*_{2,ij}$ and $y^*_{3,ij}$ and the predicted random effects (not shown here) appear roughly symmetric, this gives evidence that ignoring the folding in estimation and assuming $\boldsymbol{y} \approx |\, \boldsymbol{y} \,|$ is a valid approach. As demonstrated by Scealy and Welsh (2014a), folding will significantly bias the Kent model estimates only when there is a large amount of data on or close to the zero boundary with high variance, and this will lead to asymmetry in the unit level residual plots. Importantly, although the variability is large in this setting, the majority of the data is not too close to the

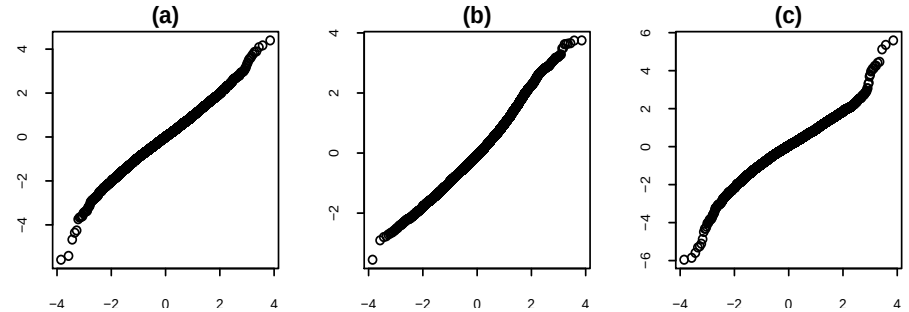

Figure 3: (a) Gaussian quantile-quantile plot of $\tilde{y}^*_{2,ij}$; (b) Gaussian quantile-quantile plot of $\tilde{y}^*_{3,ij}$; (c) Gaussian quantile-quantile plot of logratio standardised residuals.

zero boundary and the effect of folding is negligible.

6.2 Logratio model

The traditional way of treating compositional data is to apply a logratio transformation (Aitchison, 1982) and in this setting the following bivariate linear mixed model could be applied:

$$\begin{pmatrix} l_{1,ij} \\ l_{2,ij} \end{pmatrix} = \begin{pmatrix} \log\left(\frac{u_{2,ij}}{u_{1,ij}}\right) \\ \log\left(\frac{u_{3,ij}}{u_{1,ij}}\right) \end{pmatrix} = \begin{pmatrix} a_{11} + a_{12}x_{ij} + g_{1,i} + g_{2,i}z_{ij} + \upsilon_{1,ij} \\ a_{21} + a_{22}x_{ij} + g_{3,i} + g_{4,i}z_{ij} + \upsilon_{2,ij} \end{pmatrix},$$

where $\boldsymbol{g}_i = (g_{1,i}, g_{2,i}, g_{3,i}, g_{4,i})^T$ are Gaussian random effects and $\upsilon_{1,ij}$ and $\upsilon_{2,ij}$ are unit level residuals also following Gaussian distributions. The logratio transformation is undefined if either $u_{1,ij}$, $u_{2,ij}$ or $u_{3,ij}$ are zero and this occurs in a small number of cases. For simplicity we replaced all zeros with the small positive value 0.1, rescaled these observations so that $u_{1,ij} + u_{2,ij} + u_{3,ij} = 1$ and then applied the logratio transformation. We fitted two separate heteroscedastic univariate linear mixed models to the logratios $l_{1,ij}$ and $l_{2,ij}$ (we had convergence issues with running the full bivariate model) and calculated predicted unit level residuals. Plot (c) in Figure 3 is the Gaussian quantile-quantile plot of the standardised unit level residual estimates for the

$l_{1,ij}$ model showing large deviations from the Gaussian distribution. The variability in the logratios is large and the unit level residuals have heavy tails.

In small area estimation problems it is important to obtain accurate predictions of $E\left(\boldsymbol{u}_{ij}|\boldsymbol{x}_{ij},\boldsymbol{g}_i\right)$ and $E\left(\boldsymbol{u}_{ij}|\boldsymbol{x}_{ij}\right)$. For example, with a logratio model, this will involve replacing parameters with estimates and calculating E $\left(u_{1,ij}|\boldsymbol{x}_{ij},\boldsymbol{g}_i\right)$ which is equivalent to

$$\int_{-\infty}^{\infty}\int_{-\infty}^{\infty}\left(1+e^{a_{11}+a_{12}x_{ij}+g_{1,i}+g_{2,i}z_{ij}+\upsilon_{1,ij}}+e^{a_{21}+a_{22}x_{ij}+g_{3,i}+g_{4,i}z_{ij}+\upsilon_{2,ij}}\right)^{-1}\phi(\boldsymbol{\upsilon}_{ij})d\boldsymbol{\upsilon}_{ij},$$

for 8594 observations with similar calculations for $E\left(u_{2,ij}|\boldsymbol{x}_{ij},\boldsymbol{g}_i\right)$, where $\boldsymbol{\upsilon}_{ij}=(\upsilon_{1,ij},\upsilon_{2,ij})^T$ and $\phi(\boldsymbol{\upsilon}_{ij})$ is the probability density function of the bivariate Gaussian distribution. These integrals do not exist in closed form. We could use MCMC methods to approximate these integrals based on a Gaussian assumption for $\upsilon_{1,ij}$ and $\upsilon_{2,ij}$, but the Gaussian assumption is clearly violated due to the heavy tails of these residuals, making prediction difficult.

If the variability is small then under the logratio model we can approximate for example $E\left(u_{1,ij}|\boldsymbol{x}_{ij}\right)$ by

$$\left(1+e^{a_{11}+a_{12}x_{ij}}+e^{a_{21}+a_{22}x_{ij}}\right)^{-1} \tag{18}$$

and replace the parameters with estimates. Figure 4 contains a plot of $u_{1,ij}$ vs log(EXPTL) (see grey points) and estimates of $E\left(u_{1,ij}|\boldsymbol{x}_{ij}\right)$ using our square root transformation estimator (solid black line). We calculated a 95% confidence interval for $E\left(u_{1,ij}|\boldsymbol{x}_{ij}\right)$ (dashed black line) using the nonparametric bootstrap method discussed in Section 5.2. The red dotted line denoted by logratio is the estimate based on the small variance assumption (18) above. The logratio estimate is markedly different from the square root estimate when $\log(EXPTL)>7$ and is not recommended due

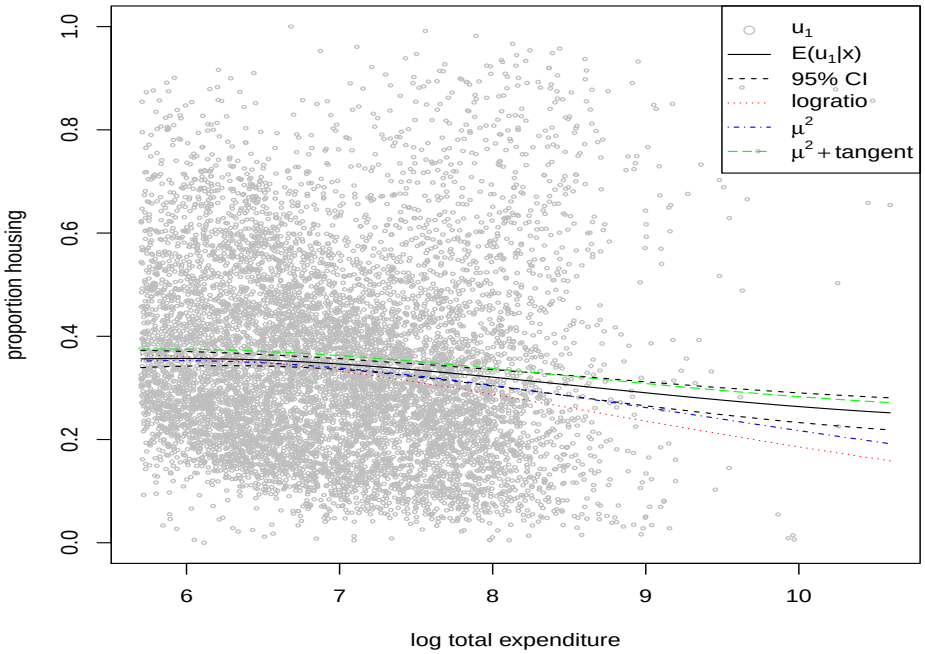

Figure 4: predictions and confidence intervals for the housing proportions

to the high variability.

6.3 Tangent space approximation

In the field of statistical shape analysis, it is common to use tangent coordinates and a Gaussian approximation in correlated data settings to simplify the analysis (e.g. Barry and Bowman, 2008). We could also use a similar approximation here after applying the square root transformation to the data. This approximation works by projecting the data from the hypersphere onto the Euclidean tangent space, and using a standard linear mixed model in the tangent space. This method will work well if the variability in the data is small (concentration is large), since in this case the surface of a sphere is well approximated by a tangent plane.

By definition $\boldsymbol{y}_{ij} = \boldsymbol{H}(\boldsymbol{x}_{ij})\boldsymbol{e}_{ij}$, where $\boldsymbol{e}_{ij}$ is the total residual including both the

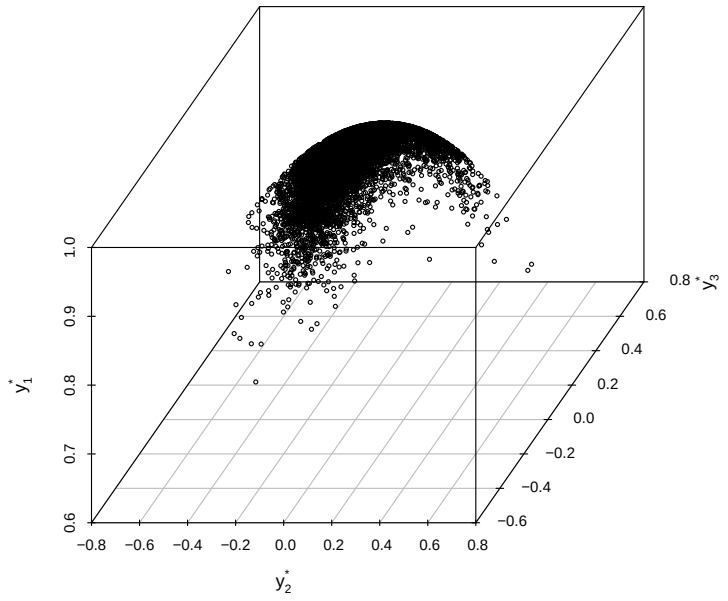

Figure 5: plot of $y^*_{1,ij}$ vs $y^*_{2,ij}$ vs $y^*_{3,ij}$ predictions (shows high curvature)

unit level errors and random effects. Consider the following decomposition

$$\begin{aligned}\boldsymbol{y}_{ij} &= e_{1,ij}\boldsymbol{\mu}(\boldsymbol{x}_{ij}) + \boldsymbol{H}^*(\boldsymbol{x}_{ij})\boldsymbol{e}_{L,ij} \\ &= \boldsymbol{\mu}(\boldsymbol{x}_{ij})\left(1+\Delta_{ij}\right) + \boldsymbol{H}^*(\boldsymbol{x}_{ij})\boldsymbol{e}_{L,ij},\end{aligned}$$

where $\Delta_{ij} = e_{1,ij} - 1$. Scealy et al. (2015) call Δ_{ij} the curvature component and $\boldsymbol{H}^*(\boldsymbol{x}_{ij})\boldsymbol{e}_{L,ij}$ the tangent space component. From the asymptotics we know $\Delta_{ij} = O_p(\kappa^{-1})$ and $\boldsymbol{e}_{L,ij} = O_p(\kappa^{-1/2})$. If we apply a tangent space approximation and ignoring the curvature component, then $E\left(\boldsymbol{u}_{ij}|\boldsymbol{x}_{ij}\right)$ is approximately equal to the diagonal elements in

$$\boldsymbol{\mu}(\boldsymbol{x}_{ij})\boldsymbol{\mu}(\boldsymbol{x}_{ij})^T + \boldsymbol{H}^*(\boldsymbol{x}_{ij})E\left(\boldsymbol{e}_{L,ij}\boldsymbol{e}_{L,ij}^T|\boldsymbol{x}_{ij}\right)\boldsymbol{H}^*(\boldsymbol{x}_{ij})^T,$$

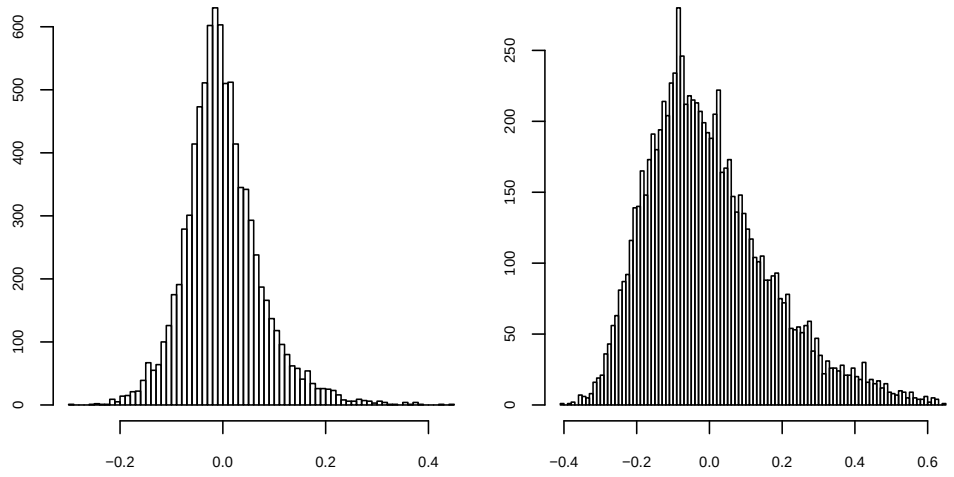

Figure 6: Histograms of the unit level residuals $w_{2,ij}$ and $w_{1,ij}$ on the original scale

or if we additionally ignore the tangent space component, we can use the approximation $\boldsymbol{\mu}(\boldsymbol{x}_{ij})^2$. In Figure 4 we include both of these approximations denoted by respectively $\boldsymbol{\mu}^2$ + tangent and $\boldsymbol{\mu}^2$, with parameters replaced by estimates from the full model fitted in Section 6.1. Both of these estimators are either borderline or markedly different from our proposed estimator. The full directional model including the curvature component is needed here because the curvature and variability is large as seen in Figure 5 and cannot be ignored. This explains geometrically why simple approximations will not work even on the square root scale.

6.4 Mixed model on original scale

Another simple alternative model is to assume that, given the random effects, the data $\boldsymbol{u}_{ij}$ on the original scale are approximately multivariate Gaussian. This approach is similar to the approach used in Militino et al. (2012), although theirs is a univariate setting. That is, assume that $\boldsymbol{u}_{ij} = E(\boldsymbol{u}_{ij}|\boldsymbol{x}_{ij}, \boldsymbol{b}_i) + \boldsymbol{e}_{ij}$, where $\boldsymbol{e}_{ij}$ has a multivariate Gaussian distribution and $\boldsymbol{b}_i$ is a vector of random effects. We can estimate unit level

residuals from this model by calculating

$$\boldsymbol{w}_{ij} = (w_{1,ij}, w_{2,ij}, w_{3,ij})^T = \boldsymbol{u}_{ij} - \hat{E}(\boldsymbol{u}_{ij}|\boldsymbol{x}_{ij}, \boldsymbol{b}_i),$$

where $\hat{E}(\boldsymbol{u}_{ij}|\boldsymbol{x}_{ij}, \boldsymbol{b}_i)$ is the prediction obtained from fitting the model in Section 6.1. Figure 6 contains histograms of $w_{2,ij}$ and $w_{1,ij}$. The distribution of the housing residuals $w_{1,ij}$ is right-skewed due to the observations being close to the zero boundary with a relatively high variance and the Gaussian assumption is violated. One advantage of the square root transformation is that it leads to approximately symmetric residual plots.

7 Simulation

We carried out a small simulation study to examine the properties of the iterative estimating equations when the approximate loglikelihood (15) is maximised to update $\boldsymbol{\Sigma}$ and $\boldsymbol{\Sigma}_b$. We considered 4 settings for $p = 3$ under the model defined in Section 2.2 (with no heteroscedasticity) and in all cases we set $\boldsymbol{x}_{ij} = 1$, $\boldsymbol{a}_1 = a_{11} = 0$ and $\boldsymbol{a}_2 = a_{21} = 0$, which implies $\boldsymbol{\mu}(\boldsymbol{x}_{ij})$ is constant. We deliberately choose $\boldsymbol{\mu}(\boldsymbol{x}_{ij})$ to be in the centre of the positive orthant so we could explore the properties of the Gaussian approximation when the folding is minimal. The cluster sizes n_i and the number of clusters m are set the same as in the example from Section 6 giving a total sample size of 8594.

We generated one set of random slope covariates z_{ij} for $i = 1, 2, \ldots, m$ and $j = 1, 2, \ldots, n_i$ independently from a standard Gaussian distribution and conditioned on these in all simulations. In all cases we also let $\boldsymbol{K}^* = \boldsymbol{I}_2$, $\boldsymbol{K}_b^* = \boldsymbol{I}_4$ and we set $\beta_2 = 0.49\kappa$ and $(\beta_{2,b}, \beta_{3,b}, \beta_{4,b})^T = \kappa_b(0.49, 0.47, 0.45)^T$. The limit conditions (9) for asymptotic normality will break down when $\beta_2 \to 0.5\kappa$ and $\beta_{2,b} \to 0.5\kappa_b$, so the

chosen shape parameters are extreme. In each setting we generated 200 samples by twice applying the Scealy and Welsh (2011) sampling algorithm. First we generated random effects from the $p = 5$ Kent distribution and then conditional on these random effects, we generated observations from the $p = 3$ folded Kent distribution (the folding step at the end guarantees that the observations are in the positive orthant).

Table 2: Bias and SE estimates for the regression coefficients

(κ, κ_b):	(200,5000)	(200,10000)	(1000,1000)	(5000,1000)
PRE a_{11}	-0.00017(0.051)	-0.00033(0.038)	0.0095(0.089)	-0.0059(0.079)
GEE a_{11}	-0.00053(0.037)	-0.0022(0.029)	0.0041(0.061)	-0.0039(0.055)
PRE a_{12}	0.0017(0.034)	0.00097(0.025)	0.0020(0.066)	-0.0019(0.065)
GEE a_{12}	-0.0013(0.024)	-0.0013(0.018)	0.0012(0.046)	0.0020(0.043)
outside	0.0035	0.0020	0.0019	0.00037
trace(Σ_b)	0.010	0.0059	0.037	0.037
trace(Σ)	0.058	0.058	0.021	0.0068

Table 3: Bias and SE estimates for $\boldsymbol{V}_{jj}^{(i)}$ and $\boldsymbol{V}_{jk}^{(i)}$ conditional on $z_{ij} = 1$ and $z_{ik} = 1$

(κ, κ_b):	(200,5000)	(200,10000)	(1000,1000)	(5000,1000)
$\boldsymbol{V}_{jj}^{(i)}[1,1]$	-2.7e-04 (0.0013)	-4.1e-04 (0.0010)	-3.2e-04 (0.0026)	-2.1e-04 (0.0026)
$\boldsymbol{V}_{jj}^{(i)}[1,2]$	-8.6e-05(5.4e-04)	2.0e-05(3.4e-04)	-6.6e-05(0.0014)	-1.5e-04(0.0014)
$\boldsymbol{V}_{jj}^{(i)}[2,2]$	-3.1e-05(3.4e-04)	1.1e-05(1.8e-04)	2.1e-05(0.0013)	-1.5e-04(0.0012)
$\boldsymbol{V}_{jk}^{(i)}[1,1]$	-7.4e-05(0.0011)	-2.9e-04(8.5e-04)	-3.6e-04(0.0027)	-2.0e-04(0.0026)
$\boldsymbol{V}_{jk}^{(i)}[1,2]$	-1.5e-04(5.1e-04)	-2.3e-05(3.2e-04)	-4.3e-05(0.0014)	-1.5e-04(0.0014)
$\boldsymbol{V}_{jk}^{(i)}[2,1]$	-8.7e-05(5.1e-04)	-1.3e-05(3.2e-04)	-7.7e-05(0.0014)	-1.7e-04(0.0014)
$\boldsymbol{V}_{jk}^{(i)}[2,2]$	-4.6e-05(3.4e-04)	-1.3e-05(1.8e-04)	1.1e-05(0.0013)	-1.5e-04(0.0012)

The 4 settings corresponded to the following 4 parameter values

$$(\kappa, \kappa_b) \in \{(200, 5000), (200, 10000), (1000, 1000), (5000, 1000)\}.$$

These choices of κ and κ_b lead to only a small amount of folding (the row labeled outside in Table 2 is the estimated proportion of observations generated outside the positive orthant before folding). The choice $\kappa = 200$ leads to a model with large unit

Table 4: Bias and SE estimates for $\boldsymbol{V}_{jj}^{(i)}$ and $\boldsymbol{V}_{jk}^{(i)}$ conditional on $z_{ij} = -3$ and $z_{ik} = 3$

(κ, κ_b):	(200,5000)	(200,10000)	(1000,1000)	(5000,1000)
$\boldsymbol{V}_{jj}^{(i)}[1,1]$	4.9e-04(0.0040)	2.2e-04 (0.0028)	0.0088(0.0093)	0.0092(0.0095)
$\boldsymbol{V}_{jj}^{(i)}[1,2]$	5.6e-05(9.7e-04)	1.5e-04 (6.3e-04)	1.2e-04 (0.0027)	-1.4e-05(0.0025)
$\boldsymbol{V}_{jj}^{(i)}[2,2]$	6.5e-05(4.3e-04)	3.3e-05(2.7e-04)	7.3e-04 (0.0016)	5.9e-04 (0.0012)
$\boldsymbol{V}_{jk}^{(i)}[1,1]$	-1.0e-04 (0.0032)	-2.0e-04 (0.0022)	-0.0026(0.0086)	-0.0043(0.0077)
$\boldsymbol{V}_{jk}^{(i)}[1,2]$	-1.7e-04(9.3e-04)	-2.3e-05(5.1e-04)	2.4e-04 (0.0027)	-2.5e-04 (0.0027)
$\boldsymbol{V}_{jk}^{(i)}[2,1]$	-2.5e-04 (0.0010)	-1.4e-04 (5.7e-04)	-2.4e-05(0.0030)	-2.7e-04 (0.0028)
$\boldsymbol{V}_{jk}^{(i)}[2,2]$	1.0e-05(3.3e-04)	3.4e-06(1.9e-04)	5.1e-04 (0.0013)	3.6e-04 (0.0012)

level variation and shape parameter estimators based on the Gaussian approximation are known to perform poorly in this case (e.g. Scealy and Welsh, 2011 Table 5; Scealy and Welsh, 2014a Table 2). In the example in Section 6 the estimate of trace($\boldsymbol{\Sigma}_b$) was 0.0057 and the estimate of trace($\boldsymbol{\Sigma}$) ranged from 0.045 to 0.068. For comparison, the rows trace($\boldsymbol{\Sigma}_b$) and trace($\boldsymbol{\Sigma}$) in Table 2 contains the average estimated values of trace($\boldsymbol{\Sigma}_b$) and trace($\boldsymbol{\Sigma}$) across the 200 simulations showing that the second simulation setting $(200, 10000)$ is similar to the example.

Monte Carlo estimates of the bias and SE (standard error) of the estimators were calculated and these are given in Tables 2, 3 and 4. In Table 2 PRE is the preliminary estimator defined by solving (11) and GEE is the estimator defined by solving (13). As expected the estimated biases for the regression coefficient estimates in Table 2 are always negligible and the GEE estimator is much more efficient than PRE. In Tables 3 and 4 $\boldsymbol{V}_{jj}^{(i)}[r,s]$ and $\boldsymbol{V}_{jk}^{(i)}[r,s]$ represent the r,sth element of the matrices $\boldsymbol{V}_{jj}^{(i)}$ and $\boldsymbol{V}_{jk}^{(i)}$ respectively. The true values needed in the bias calculations were accurately approximated by simulating a large sample of size 100000 from the true model conditional on the covariates $z_{ij} = 1$ and $z_{ik} = 1$ (Table 3) or $z_{ij} = -3$ and $z_{ik} = 3$ (Table 4). Most of the estimated biases in Tables 3 and 4 are also negligible. When $\kappa_b = 1000$ and $z_{ij} = -3$ and $z_{ik} = 3$ the estimator of $\boldsymbol{V}_{jj}^{(i)}[1,1]$ is a little

more biased, but not significantly so since | Bias | /SE < 1. Therefore the Gaussian approximation for the variance components works extremely well even for moderately large values of $\mathbf{\Sigma}_b$.

8 Conclusion

We introduced a new class of mixed model for directional data based on Kent distributions. A valuable property of these models is that the marginal mean directions are available in a closed form and a convenient quasi-likelihood moment based method can be used for estimation. The new directional mixed model was successfully applied to analyse a large compositional household expenditure survey dataset with observations clustered in small domains and we showed that the estimates are interpretable. The simulation study confirmed that the proposed estimators have low bias in typical small area estimation contexts. In comparison, the logratio method and other simple approaches such as the tangent space approximation do not work well due to the high variability in the data. Prediction is also difficult under the logratio model due to the violation of the Gaussian assumption on the logratio scale. The Kent mixed effect models assumed that the unit level residuals are independent and future work will involve adapting the models to handle spatial and other more general correlation structures. Other future work will also involve developing robust estimators to help deal with potential outliers and other small model departures.

Acknowledgments

This research was supported by an Australian Research Council discovery project grant. We thank two referees and an Associate Editor for their reviews which have lead to an improved paper.

A

In all expectations below we are implicitly conditioning on $\boldsymbol{x}$. Let $\boldsymbol{G} = \boldsymbol{A}\boldsymbol{H}^{*T}\boldsymbol{y}$ and $\boldsymbol{G}^* = \frac{\partial \boldsymbol{\mu}}{\partial \boldsymbol{a}}\boldsymbol{H}^*\boldsymbol{V}^{-1}\boldsymbol{H}^{*T}\boldsymbol{y}$. Then

$$\mathrm{E}(\boldsymbol{G}\boldsymbol{G}^{*T}) = \boldsymbol{A}\,\mathrm{E}\left(\boldsymbol{H}^{*T}\boldsymbol{y}\boldsymbol{y}^T\boldsymbol{H}^*\right)\boldsymbol{V}^{-1}\boldsymbol{H}^{*T}\left(\frac{\partial \boldsymbol{\mu}}{\partial \boldsymbol{a}}\right)^T = \boldsymbol{A}\boldsymbol{H}^{*T}\left(\frac{\partial \boldsymbol{\mu}}{\partial \boldsymbol{a}}\right)^T.$$

Using the orthogonality result $\boldsymbol{H}^{*T}\boldsymbol{\mu} = \boldsymbol{0}$, it follows that for the random intercept model defined in Section 2.1

$$\mathrm{E}\left(\frac{\partial \boldsymbol{G}}{\partial a_j}\right) = \boldsymbol{A}\frac{\partial \boldsymbol{H}^{*T}}{\partial a_j}\boldsymbol{\mu}\,\mathrm{E}(b_{1,1})\psi$$

for any component a_j of $\boldsymbol{a}$. The orthogonality result also implies

$$\frac{\partial(\boldsymbol{H}^{*T}\boldsymbol{\mu})}{\partial a_j} = \boldsymbol{0} = \frac{\partial \boldsymbol{H}^{*T}}{\partial a_j}\boldsymbol{\mu} + \boldsymbol{H}^{*T}\frac{\partial \boldsymbol{\mu}}{\partial a_j}$$

and hence

$$\mathrm{E}\left(\frac{\partial \boldsymbol{G}}{\partial \boldsymbol{a}}\right) = -\,\mathrm{E}(b_{1,1})\psi\boldsymbol{A}\boldsymbol{H}^{*T}\left(\frac{\partial \boldsymbol{\mu}}{\partial \boldsymbol{a}}\right)^T.$$

It follows that

$$\left(\mathrm{E}\left(\frac{\partial \boldsymbol{G}}{\partial \boldsymbol{a}}\right)\right)^{-1}\mathrm{E}(\boldsymbol{G}\boldsymbol{G}^{*T})$$

is constant and $\boldsymbol{G}^*$ is optimal by Theorem 2.1 of Heyde (1997). In the case of the random slope model defined in Section 2.2,

$$\mathrm{E}\left(\frac{\partial \boldsymbol{G}}{\partial a_j}\right) = \boldsymbol{A}\frac{\partial \boldsymbol{H}^{*T}}{\partial a_j}\boldsymbol{\mu}(1 + O(\kappa_b^{-1}))\psi \approx \boldsymbol{A}\frac{\partial \boldsymbol{H}^{*T}}{\partial a_j}\boldsymbol{\mu}\psi$$

and therefore $\boldsymbol{G}^*$ is approximately optimal for the random slope model.

References

Australian Bureau of Statistics. (2012). *Household Expenditure Survey and Survey of Income and Housing, User Guide, Australia, 2009-10.* Catalogue number 6503.0, ABS, Canberra.

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

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

Barry, S. J. E. and Bowman, A. W. (2008). Linear mixed models for longitudinal shape data with applications to facial modeling. *Biostatistics* **9** 555–565.

Chatterjee, S. and Bose, A. (2005). Generalized bootstrap for estimating equations. *The Annals of Statistics* **33** 414–436.

Field, C. A., Pang, Z. and Welsh, A. H. (2010). Bootstrapping robust estimates for clustered data *Journal of the American Statistical Association.* **105** 1606–1616.

Haddou, M., Rivest, L. and Pierrynowski, M. (2010). A nonlinear mixed effects directional model for the estimation of the rotation axes of the human ankle. *The Annals of Applied Statistics* **4** 1892–1912.

Heyde, C. C. (1997). *Quasi-Likelihood And Its Application: A General Approach to Optimal Parameter Estimation*. Springer, New York.

Jiang, J. and Lahiri, P. (2006). Mixed model prediction and small area estimation. *TEST* **15** 1–96 .

Jiang, J. (2007). *Linear and generalized linear mixed models and their applications*. Springer, New York.

Jiang, J., Luan, Y. and Wang, Y (2007). Iterative estimating equations: linear convergence and asymptotic properties. *The Annals of Statistics* **35** 2233–2260.

Kent, J. T. (1982). The Fisher-Bingham distribution on the sphere. *Journal of the*

Royal Statistical Society, Series B **44** 71–80.

Mardia, K. V. and Jupp, P. E. (2000). *Directional Statistics*. Wiley, Chichester.

McCullagh, P. and Nelder, J. A. (1989). *Generalized Linear Models*, Second Edition. Chapman and Hall, London.

Militino, A. F., Goicoa, T. and Ugarte, M. D. (2012). Estimating the percentage of food expenditure in small areas using bias-corrected P-spline based estimators. *Computational Statistics & Data Analysis* **56** 2934–2948.

Molina, I., Saei, A. and Lombarda, M. J. (2007). Small area estimates of labour force participation under a multinomial logit mixed model. *Journal of the Royal Statistical Society Series A* **170** 975–1000 .

Pfeffermann, D. (2013). New important developments in small area estimation. *Statistical Science* **28** 40–68.

Pinheiro, J. C. and Bates, D. M. (2000). *Mixed-effects models in S and S-PLUS*. Springer, New York.

Rao, J. N. K. (2003). *Small area estimation*. John Wiley and Sons, Hoboken.

Samanta, M. and Welsh, A. H. (2013). Bootstrapping for highly unbalanced clustered data. *Computational Statistics & Data Analysis* **59** 70–81.

Scealy, J. L. (2010). *Modelling techniques for compositional data using distributions defined on the hypersphere*. Thesis (Ph.D.)- Australian National University.

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* **73** 351–375.

Scealy, J. L. and Welsh, A. H. (2014a). Fitting Kent models to compositional data with small concentration. *Statistics and Computing* **24** 165–179

Scealy, J. L. and Welsh, A. H. (2014b). Colours and cocktails: compositional data analysis 2013 Lancaster lecture. *Australian & New Zealand Journal of Statis-*

tics **56** 145–169

Scealy, J. L., Caritat, P. de, Grunsky, E. C., Tsagris, M. T., Welsh, A. H. (2015). Robust principal component analysis for power transformed compositional data. *Journal of the American Statistical Association* **110** 136–148.

Stephens, M. A. (1982). Use of the von Mises distribution to analyse continuous proportions. *Biometrika* **69** 197–203.