Small area estimation of expenditure proportions

Janice Scealy \textsuperscript{1} and Alan Welsh \textsuperscript{2}

\textsuperscript{1}Research School of Finance, Actuarial Studies and Statistics, ANU

\textsuperscript{2}Mathematical Sciences Institute, ANU

July 2017
Introduction

• We develop a new mixed effects model for compositional data.
• Compositional data are vectors $u$ defined on the unit simplex

$$\Delta^{p-1} = \left\{ (u_1, u_2, \ldots, u_p)^T \in \mathbb{R}^p : u_i \geq 0, \sum_{i=1}^{p} u_i = 1 \right\}.$$

• It is also possible for the compositional data to be correlated due to the grouping of the observations within small domains.
• The square root transformation: $y = (\sqrt{u_1}, \sqrt{u_2}, \ldots, \sqrt{u_p})^T$.
• Transforms $u$ onto the surface of the $p-1$ dimensional hypersphere $S^{p-1} = \{ y \in \mathbb{R}^p : \|y\| = 1 \}$.
• This opens the possibility of using distributions for directional data to model compositional data.
• Scealy and Welsh (2011) proposed the Kent distribution for modelling $y$. 
The additive Kent regression model

- The Kent distribution for $p \geq 3$ is defined by the density:

$$f(y) = c(\kappa, \beta)^{-1} \exp[\kappa \mu^T y + y^T \Gamma D_c \Gamma^T y], \quad y \in S^{p-1},$$

where $D_c = \text{diag} \left(0, \beta^T, \sum_{m=2}^{p-1} \beta_m \right)$ and $\Gamma = (\mu, \gamma_2, \ldots, \gamma_p)$.

- **Parameters**: $\Gamma$ is an orthogonal matrix (location parameters). $\kappa$ and $\beta = (\beta_2, \beta_3, \ldots, \beta_{p-1})^T$ are shape parameters.

- **Regression Model**: $\Gamma = \Gamma(x)$ is a function of covariates $x \in \mathbb{R}^q$, and the shape parameters are constant. Let

$$\Gamma(x) = \begin{pmatrix} \mu_1(x) \\ \mu_L(x) \end{pmatrix} \begin{pmatrix} \mu_L(x)^T \\ 1 + \mu_1(x) \mu_L(x) \mu_L(x)^T - I_{p-1} \end{pmatrix} \begin{pmatrix} 1 & 0^T \\ 0 & K^* \end{pmatrix},$$

where $\mu(x) = (\mu_1(x), \mu_2(x), \ldots, \mu_p(x))^T = (\mu_1(x), \mu_L(x)^T)^T$.

Then define

$$\mu_k(x) = \begin{cases} \left(1 + \sum_{j=1}^{p-1} \exp \left( a_j^T x \right) \right)^{-\frac{1}{2}} & k = 1 \\ \exp \left( \frac{a_{k-1}^T x}{2} \right) \left(1 + \sum_{j=1}^{p-1} \exp \left( a_j^T x \right) \right)^{-\frac{1}{2}} & k = 2, 3, \ldots, p, \end{cases}$$
Kent mixed effects model

• The regression model can be summarised as

\[ y_i \mid x_i \sim Kent(\Gamma(x_i), \kappa, \beta) \quad i = 1, 2, \ldots, n. \]

• Now instead assume the sample consists of

\[ y_{ij} = \sqrt{u_{ij}} \quad \text{and} \quad x_{ij}, \quad j = 1, 2, \ldots, n_i, \quad i = 1, 2, \ldots, m. \]

• Let \( b_i \) be a vector of random effects in cluster \( i \). We assume:

\[ b_i \sim Kent(\Gamma_b, \kappa_b, \beta_b) \quad i = 1, 2, \ldots, m, \quad \text{and} \quad \Gamma_b = \left( \begin{array}{c} 1 \\ 0 \\ K_b^* \end{array} \right). \]

• The mixed effects model is:

\[ y_{ij} \mid x_{ij}, b_i \sim Kent(\Gamma(x_{ij}, b_i), \kappa, \beta), \quad j = 1, 2, \ldots, n_i, \quad i = 1, 2, \ldots, m, \]
The location matrix $\Gamma$

$$\Gamma(x_{ij}, b_i) = H(x_{ij})R(b_i)\left( \begin{array}{cc} 1 & 0^T \\ 0 & K^* \end{array} \right),$$

where

$$H(x_{ij}) = \left( \begin{array}{c} \mu_1(x_{ij}) \\ \mu_L(x_{ij}) \end{array} \right) - \frac{1}{1 + \mu_1(x_{ij})} \left( \begin{array}{c} \mu_L(x_{ij})^T \\ \mu_L(x_{ij})^T \end{array} \right) - I_{p-1},$$

$$R(b_i) = \left( \begin{array}{cc} b_{1,i} & -b_{L,i}^T \\ b_{L,i} & I_{p-1} - \frac{1}{1+b_{1,i}} b_{L,i} b_{L,i}^T \end{array} \right),$$

where $b_i = (b_{1,i}, b_{2,i}, \ldots, b_{p,i})^T = (b_{1,i}, b_{L,i}^T)^T$.

- $R(b_i)$ is a rotation matrix.

- Units in the same cluster with the same covariates will tend to have more similar locations than units in different clusters.
Random slope model

- We can also incorporate random slopes in the model, allowing more flexibility when modelling the between cluster variation.
- Let \( \{z_{ij} \in \mathbb{R} : j = 1, 2, \ldots, n_i; i = 1, 2, \ldots, m\} \) be a set of standardised random variables.
- Define \( b_i \in S^{2p-2} \) to have a standardised Kent distribution with \( 2p - 2 \) shape parameters and with orthogonal location matrix \( K_b^{*} \) of size \( (2p - 2) \times (2p - 2) \).
- Let \( \xi_{ij} = (\xi_{1,ij}, \xi_{2,ij}, \ldots, \xi_{p,ij})^T \), where \( \xi_{1,ij} = b_{1,i} \) and \( \xi_{r,ij} = b_{r,i} + z_{ij} b_{p+r-1,i} \) for \( r = 2, 3, \ldots, p \).
- Project onto \( S^{p-1} \): \( \xi_{ij}^* = (\xi_{1,ij}^*, \xi_{2,ij}^*, \ldots, \xi_{p,ij}^*)^T = \xi_{ij}/ \| \xi_{ij} \| \).
- Replace \( \Gamma(x_{ij}, b_i) \) in the conditional model by

\[
\Gamma(x_{ij}, b_i) = H(x_{ij})R(\xi_{ij}^*) \begin{pmatrix} 1 & 0^T \\ 0 & K^* \end{pmatrix},
\]

where the rotation matrix \( R(.) \) has the same form as before.
Moments

Conditional mean:

\[ E(y_{ij}|x_{ij}, b_i) \propto H(x_{ij}) \xi_{ij}^*, \quad j = 1, 2, \ldots, n_i, \quad i = 1, 2, \ldots, m. \]

Marginal mean:

\[ E(y_{ij}|x_{ij}) \propto \mu(x_{ij}), \quad j = 1, 2, \ldots, n_i, \quad i = 1, 2, \ldots, m. \]

Second order marginal moments:

\[
E(y_{ij}y_{ij}^T|x_{ij}) = H(x_{ij})\text{block diag } \left( 1 - \text{trace} \left( V_{jj}^{(i)} \right), V_{jj}^{(i)} \right) H(x_{ij})^T
\]

(1)

\[
E(y_{ij}y_{ik}^T|x_{ij}, x_{ik}) = H(x_{ij})\text{block diag } \left( v_{jk}^{(i)}, V_{jk}^{(i)} \right) H(x_{ik})^T \quad j \neq k
\]

\[
E(y_{ij}y_{rk}^T|x_{ij}, x_{rk}) \propto \mu(x_{ij})\mu(x_{rk})^T \quad i \neq r.
\]

\(E(u_{ij}|x_{ij})\) is obtained from the diagonal elements of (1) since \(u_{ij} = y_{ij}^2\).

\(V_{jj}^{(i)}, V_{jk}^{(i)}\) and \(v_{jk}^{(i)}\) are functions of \(z_{ij}, z_{ik}, K^*, K_b^*\) and the shape parameters.
Estimation and Inference

- ML is difficult, instead we use quasi-likelihood estimation and a normal approximation:

1. Estimate $a$ given $V_i$ by solving $\sum_{i=1}^{m} \frac{\partial \mu_i}{\partial a} H_i^* V_i^{-1} H_i^* y_i = 0$.
2. Normal approximation (valid when $\kappa$ and $\kappa_b$ are large):

$$H_i^* y_i \sim N_{n_i(p-1)}(0, V_i) , \quad i = 1, 2, \ldots, m,$$

$$V_{jj}^{(i)} \approx \Sigma + A_{ij} \Sigma_b A_{ij}^T \quad \text{and} \quad V_{jk}^{(i)} \approx A_{ij} \Sigma_b A_{ik}^T \quad (j \neq k),$$

where $A_{ij} = (I_{p-1} \quad z_{ij}I_{p-1})$ or $A_{ij} = I_{p-1}$. Given $a$, maximise the approximate normal loglikelihood with respect to $\Sigma$ and $\Sigma_b$.

- To obtain $\hat{a}$, $\hat{V}_{jj}^{(i)}$ and $\hat{V}_{jk}^{(i)}$ iterate between steps (1) and (2) a fixed number of times or until convergence (see Jiang et al. (2007)).

- From simulations we confirmed that $\hat{V}_{jj}^{(i)}$ and $\hat{V}_{jk}^{(i)}$ are approximately unbiased even when $\kappa$ and $\kappa_b$ are not large.

- For inference we apply 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).
Application

- Militino et al. (2012) fit a P-spline model to clustered compositional food expenditure data from the 2006 Spanish Household Budget Survey.
- We analyse the 2009-10 Australian HES CURF.
- There are $i = 1, 2\ldots, 156$ small domains of interest defined as the cross-classification between the variables State $\times$ Area of Usual Residence $\times$ Life Cycle Group. Here $n = \sum_{i=1}^{156} n_i = 8594$.
- Let $y_{ij} = (\sqrt{u_{1,ij}}, \sqrt{u_{2,ij}}, \sqrt{u_{3,ij}})^T$ where $u_{ij}$ represents expenditure proportions in 3 classes: food, housing and other.
- $x_{ij} = (1, x_{ij})^T = \left(1, \frac{\log(EXPTL_{ij}) - m^*}{s^*}\right)^T$ and $z_{ij} = x_{ij}$.
- Regression coefficients: $a_1 = (a_{11}, a_{12})^T$ and $a_2 = (a_{21}, a_{22})^T$.
- $\Sigma_b$ is a $4 \times 4$ symmetric matrix with $r, s$th element $\tau_{rs}$.

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

where $v_{ij} = c_2^{-1} \log(EXPTL_{ij})$ and $c_2 = 10.59663$. 
Comparison to other methods

Logratio method:

- Fit a bivariate LMM to \( \left( \log \frac{u_{2,ij}}{u_{1,ij}}, \log \frac{u_{3,ij}}{u_{1,ij}} \right)^T \).
- \( E(u_{1,ij}|x_{ij}) \) not in closed form and relies on Gaussian assumption.
- If variability small \( E(u_{1,ij}|x_{ij}) \approx (1 + e^{a_{11} + a_{12}x_{ij}} + e^{a_{21} + a_{22}x_{ij}})^{-1} \).
- Residual plots show high variability and heavy tails.

Mixed model on original scale:

- \( u_{ij} = E(u_{ij}|x_{ij}, b_i) + e_{ij} \)
- Could get negative predictions.
- The residuals are skewed due to the boundaries of the simplex.

Tangent space approximation (on the square root scale):

- \( y_{ij} = \mu(x_{ij})(1 + \Delta_{ij}) + H^*(x_{ij})(e_{2,ij}, e_{3,ij}, \ldots, e_{p,ij})^T \),
- If the variability is small \( E(u_{ij}|x_{ij}) \approx \text{diagonal elements in } \mu(x_{ij})\mu(x_{ij})^T + H^*(x_{ij})E(e_L,ij e_L,ij^T|x_{ij})H^T(x_{ij}) \)
  or \( \mu(x_{ij})\mu(x_{ij})^T \).
Figure: predictions and confidence intervals for the housing proportions

- $u_1$
- $E(u_1|x)$
- 95% CI
- logratio
- $\mu^2$
- $\mu^2 + \text{tangent}$

proportion housing

log total expenditure
Figure: predictions and confidence intervals for the food proportions.
Conclusion

- A new directional mixed effects model was successfully applied to analyse a large HES dataset with observations clustered in small domains.
- The logratio method and other simple models do not work well here due to the high variability.
- The simulation study confirmed that the proposed estimators have low bias in typical cases (not shown here).

Robustness:

- In future work we will develop new robust estimators to help deal with potential outliers and other small model departures.
- The sample space is compact (bounded), so our estimators here are reasonably robust already.
- However outliers are possible if the data are concentrated.
- Plan to develop new SB-robust (standardised bias robust) estimators for spherical data.
- One idea is to define a new set of M-estimators using a similar idea to Cantoni and Ronchetti (2001).
Selected References


