Transforming Compositional Data for Analysis

Cause of Death Mortality Data

June 2023

Outline

1. Introduction and literature
2. How to forecast mortality using CODA?
3. Ways to transform compositional data for analysis

What is compositional data

- Non-negative data summing to a fixed constant
- Scale invariance: ratios between components of a composition are important, unaffected by the set of components chosen
- Sub-compositional coherence: relativities in subsets remain consistent

Mortality by cause as a composition

Consider death counts by age and cause as compositional, density of deaths subject to unit constraint.

Denote actual death counts as $D_{t,x,i}$, where $t =$ year of death, x $=$ age band per data, $i =$ cause per data.

Cause-specific life table deaths:

$$d_{t,x,i} = \frac{D_{t,x,i}}{D_t}.$$

Denote the vector of compositional deaths as $\mathbf{d_t}$, where each row is a composition of deaths by age and cause for year t.

Ways to analyse compositional data

1. Ignore the compositional constraint and apply standard multivariate statistical analysis: raw data analysis (RDA)
2. Transform the compositional data using log-ratios, maintaining the properties of compositional data: log-ratio analysis (LRA)

Literature

- Cause-specific mortality forecast using life table deaths and density of deaths in a CODA framework[1]
- Cause-specific mortality modelling joint and individual variation between causes[2]
- Not limited to understanding mortality by cause: subnational levels[3], socio-economic groups[4], multi-state life tables[5]

1 Oeppen, 2008
2 Kjaergaard et al., 2019
3 Bergeron-Boucher et al., 2017
4 Kjaergaard et al., 2020
5 Bergeron-Boucher et al., 2022

How to forecast mortality for compositional data?

1. Transform compositional data from constrained space (simplex) to the unconstrained space
2. Apply standard statistical techniques to forecast mortality.
3. Transform the estimated forecast deaths back to the compositional space

Today, we focus on step 1 (and 3).

Transformations

- Two log-ratio transformations:
  - Centred Log-Ratio
  - Isometric Log-Ratio
- Challenges with zeroes in compositional data
- Flexibility through α-transformation[6]

[6] Tsagris et al., 2011

Centred Log-Ratio Transformation

$$\mathsf{CLR}(d_{t,x,i}) = w^0(d_{t,x,i}) = \ln \frac{d_{t,x,i}}{(\prod_{t,x,i} d_{t,x,i})^{1/XI}}; \qquad XI = 1, \ldots, X.$$

For the compositional vector $\mathbf{d_t}$, let $\tilde{\mathbf{d}}_\mathbf{t}$ represent the centred vector. The CLR is:

$$\mathsf{CLR}(\mathbf{d_t}) = w^0(\mathbf{d_t}) = \ln(\tilde{\mathbf{d}}_\mathbf{t}).$$

Isometric Log-Ratio Transformation

Transforms compositional data to the Euclidean space without zero-sum constraint.

More complex than CLR: ILR is the log of geometric means of two subsets of parts.

ILR is equivalent to left multiplication of CLR by Helmert sub-matrix ($\mathbf{H}$).

Still not able to work with zeroes in the data, but we have removed one constraint (zero-sum).

Isometric Log-Ratio Transformation

Denote the ILR of the compositional vector $\mathbf{d_t}$ as $z^0(\mathbf{d_t})$:

$$z^0(\mathbf{d_t}) = \mathbf{H}w^0(\mathbf{d_t}),$$

where $w^0(\mathbf{d_t})$ is the CLR transformation of the compositional vector $\mathbf{d_t}$.

Alpha Transformation

Flexibility depending on data: a Box-Cox transformation applied to the ratios of components, the selection of α determines the transformation.

Define the α-transformation as:

$$z^{\alpha}(\tilde{\mathbf{d}}_t) = \mathbf{H} w^{\alpha}(\tilde{\mathbf{d}}_t),$$

where $\tilde{\mathbf{d}}_t$ is the vector of centred death densities.

$$w^{\alpha}(\tilde{\mathbf{d}}_t) = \begin{cases} \ln(\tilde{\mathbf{d}}_t) & \alpha = 0 \\ (XI) \times \frac{\tilde{\mathbf{d}}_t^{\alpha} - 1}{\alpha} & \alpha \neq 0. \end{cases}$$

Alpha Transformation

The α-transformation maps compositional data back to the real space.

It has neither the sum constraint, nor is constrained with zeroes in the data.

When there are not a lot of zeroes in the data, the optimal α converges to zero, and the transformation converges to the ILR.

Preliminary Results

Table 1: RMSE: LC, LC-LRA, LC-α

Model and Transformation	RMSE
LC	0.0090
LC CODA ($\alpha = 0$, LRA)	0.0059
LC CODA ($\alpha = 0.1$)	0.0216
LC CODA ($\alpha = 0.5$)	0.0199
LC CODA ($\alpha = 1$)	0.0198

Thank you

Email: zhe.michelle.dong@gmail.com