Linear Mixed-Effects Models for Longitudinal Compositional Data

Zhichao Wang [1] Huiwen Wang [12] Shanshan Wang [12]

[1]School of Economics and Management, Beihang University, China

[2]Beijing Key Laboratory of Emergence Support Simulation Technologies for City Operations, China

January 22, 2018

Contents

Contents

Compositional data (CoDa)

Compositional data (or compositions) describe the intrinsic structure of a whole, which only carry relative information by proportion or percentage. Any D-part composition can be expressed as a column vector with D **positive components**, $\mathbf{x} = [x_1, x_2, \cdots, x_D]'$, satisfying the **constant-sum constraint**. That is,

$$0 < x_d < 1 \quad (d = 1, 2, \cdots, D)$$

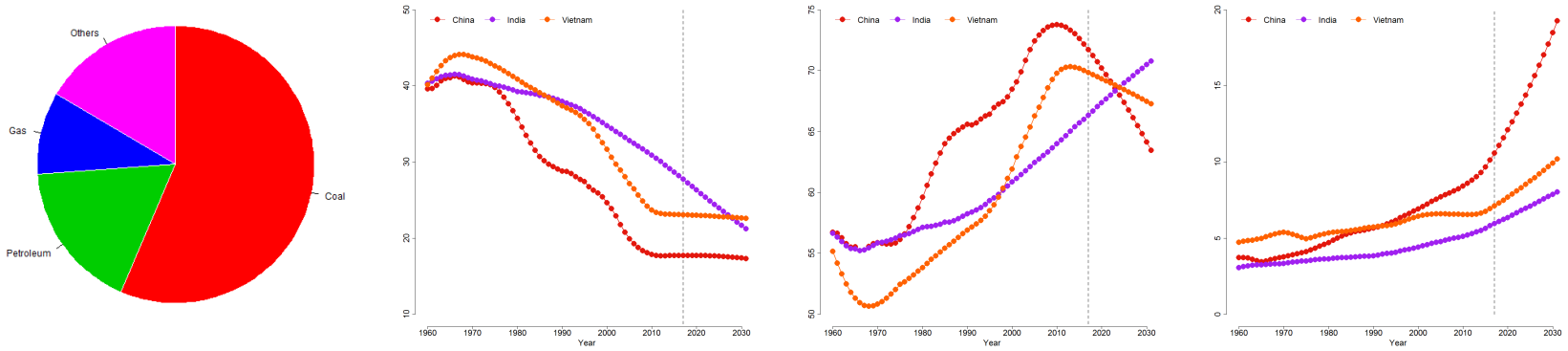
and

$$\sum_{d=1}^{D} x_d = 1.$$

All D-part compositions constitute the D-part Simplex space, denoted by S^D.

Compositional data (CoDa)

In economic data analysis, composition data are introduced to express the economic indicators containing structural information. For instance, the energy consumption structure, industrial structure and population structure. These structural economic indicators may have some correlation. To qualify this correlation, we consider regression models.



(a) (b) (c) (d)

Figure: (a) shows the forecasted energy consumption structure of China by 2020(b)-(d) show the trend of the percentages of the young, the middle and the old people of China (red), India (orange) and Vietnam (purple), respectively, since 1960.

Longitudinal compositional data

In many cases, compositional data observations also show the significant feature of longitudinal data. Measurements of structural economic indicators were taken from individuals (regions or countries) through time, which results in the **heterogeneity** and the **dependency** in the population. We call them **longitudinal compositional data**.

As will be discussed in application, the data table we used can be summarized as

"Area $\times$ Year $\times$ Indicators".

- Inertia of economic activities: samples from the same individual are highly correlated.
- Difference between areas: correlations of variables vary in different areas.
- **Samples will not be independent!**

Regression methods for CoDa

Owing to the constraints, traditional statistical methods **are not applicable directly** for compositional data. For the regression models with both compositional response and covariates, the existing research results are relatively few.

- Wang et al. (2013) regarded the whole compositional data variable as a unit.
- Chen et al. (2016) considered every part of all the compositional data variables.

The existing regression methods are conducted under the independent identically distributed assumption, which fail for longitudinal compositional data.

Hence the study for longitudinal compositional data modeling is of necessity.

Linear mixed-effects models (LMM)

The linear mixed-effects model is a common approach to deal with longitudinal data, which considers the individual difference by introducing a series of random effect.

For traditional real variables, linear mixed-effects models are expressed as

$$y_{ij} = \mathbf{x}_{ij}'\boldsymbol{\beta} + \mathbf{z}_{ij}'\mathbf{b}_i + \varepsilon_{ij},$$

where $i = 1, 2, \cdots, N$, $j|i = 1, 2, \cdots, n_i$ and

$$\begin{aligned} \mathbf{x}_{ij} &= (x_{ij1}, x_{ij2}, \cdots, x_{ijp})' \in \mathbb{R}^p, \\ \mathbf{z}_{ij} &= (z_{ij1}, z_{ij2}, \cdots, z_{ijq})' \in \mathbb{R}^q. \end{aligned}$$

$\mathbf{x}_{ij}'\boldsymbol{\beta}$ and $\mathbf{z}_{ij}'\mathbf{b}_i$ are referred to as **fixed effect** and **random effect**, respectively. Both $\mathbf{x}_{ij}$ and $\mathbf{z}_{ij}$ are composed of some covariates. Coefficients of random effect $\mathbf{b}_i$ consist of a random vector, associated with the individual i.

LMM for longitudinal compositional data (CoLMM)

Here we attempt to expand traditional LMM approach to longitudinal compositional data analysis. In advanced, we propose the following model:

$$\mathbf{y}_{ij} = \mathbf{x}_{ij}' \odot \boldsymbol{\beta} \oplus \mathbf{z}_{ij}' \odot \mathbf{b}_i \oplus \boldsymbol{\varepsilon}_{ij}, \tag{1}$$

where $i = 1, 2, \cdots, N$, $j | i = 1, 2, \cdots, n_i$ and

$$\begin{aligned} \mathbf{x}_{ij} &= [\mathbf{x}_{ij1}, \mathbf{x}_{ij2}, \cdots, \mathbf{x}_{ijp}]' \in S_p^D, \\ \mathbf{z}_{ij} &= [\mathbf{z}_{ij1}, \mathbf{z}_{ij2}, \cdots, \mathbf{z}_{ijq}]' \in S_q^D. \end{aligned}$$

Analogously, we say $\mathbf{x}_{ij}' \odot \boldsymbol{\beta}$ and $\mathbf{z}_{ij}' \odot \mathbf{b}_i$ are **fixed effect** and **random effect**, respectively. Both $\mathbf{x}_{ij}$ and $\mathbf{z}_{ij}$ are composed of some covariates. Coefficients of random effect $\mathbf{b}_i$ consist of a random vector, associated with the individual i.

Our task is to estimate $\boldsymbol{\beta}$ and those determine $\mathbf{b}_i$ and $\boldsymbol{\varepsilon}_{ij}$.

Contents

Aitchison geometry

Note that Euclidean geometry dose not apply to the Simplex S^D.

In Aitchison geometry, S^D is induced to be a vector space. For any compositions $\mathbf{x} = [x_1, x_2, \cdots, x_D]'$ and $\mathbf{y} = [y_1, y_2, \cdots, y_D]' \in S^D$, $\mathbf{z} = (z_1, z_2, \cdots, z_D)' \in \mathbb{R}_+^D$, and $\alpha \in \mathbb{R}$, three main operators are defined as

- **Closure ($\mathcal{C}$):**

$$\mathcal{C}(\mathbf{z}) = \left[\frac{z_1}{\sum_{d=1}^D z_d}, \frac{z_2}{\sum_{d=1}^D z_d}, \cdots, \frac{z_D}{\sum_{d=1}^D z_d}\right]' \in S^D$$

- **Perturbation ($\oplus$):**

$$\mathbf{x} \oplus \mathbf{y} = \mathcal{C}(x_1 y_1, x_2 y_2, \cdots, x_D y_D)'$$

- **Powering ($\odot$):**

$$\alpha \odot \mathbf{x} = \mathcal{C}(x_1^\alpha, x_2^\alpha, \cdots, x_D^\alpha)'$$

Isomorphism logratio (ilr) transformation

The isomorphism logratio (ilr) transformation is widely introduced to remove the constraints of compositional data. Given a contact matrix $\mathbf{\Phi} \in \mathbb{R}^{(D-1)\times D}$, $\mathbf{x} \in S^D$ is transformed to $\mathbb{R}^{D-1}$ by the ilr transformation. That is,

$$\mathrm{ilr}(\mathbf{x}) = \mathbf{x}^* = \mathbf{\Phi} \cdot \log(\mathbf{x}).$$

The contact matrix $\mathbf{\Phi}$ is **one-to-one corresponding to** a certain ilr transformation, and satisfies that $\mathbf{\Phi}'\mathbf{\Phi} = \mathbf{I}_{D-1}$. One of the common contact matrixes is

$$\mathbf{\Phi} = \begin{pmatrix} \frac{1}{\sqrt{D(D-1)}} & \frac{1}{\sqrt{D(D-1)}} & \cdots & \frac{1}{\sqrt{D(D-1)}} & -\sqrt{\frac{D-1}{D}} \\ \frac{1}{\sqrt{(D-1)(D-2)}} & \frac{1}{\sqrt{(D-1)(D-2)}} & \cdots & \sqrt{\frac{D-2}{D-1}} & 0 \\ \vdots & \vdots & \vdots & \vdots & \vdots \\ \frac{1}{\sqrt{2}} & -\sqrt{\frac{1}{2}} & \cdots & \cdots & 0 \end{pmatrix}.$$

Matrix form for CoDa

Multiple D-part Simplex space S_p^D is defined as the Cartesian product of p Simplex S^D, i.e.,

$$S_p^D = \left\{ \mathbf{x} = \left[\mathbf{x}_1', \mathbf{x}_2', \cdots, \mathbf{x}_p' \right]' : \mathbf{x}_k \in S^D, k = 1, 2, \cdots, p \right\}.$$

Obviously$S_1^D = S^D$. Elements of S_p^D are called p-dimensional D-part compositional data.

Pile up n p-dimensional compositions $\mathbf{x}_i$ $(i = 1, 2, \cdots, n)$ by row, then the composition matrix can be expressed in matrix form as

$$\mathbf{x} = \begin{pmatrix} \mathbf{x}_1' \\ \mathbf{x}_2' \\ \vdots \\ \mathbf{x}_n' \end{pmatrix} = \begin{pmatrix} \mathbf{x}_{11}' & \mathbf{x}_{12}' & \cdots & \mathbf{x}_{1p}' \\ \mathbf{x}_{21}' & \mathbf{x}_{22}' & \cdots & \mathbf{x}_{2p}' \\ \vdots & \vdots & \ddots & \vdots \\ \mathbf{x}_{n1}' & \mathbf{x}_{n2}' & \cdots & \mathbf{x}_{np}' \end{pmatrix}.$$

Matrix form for CoDa

Using the aforementioned matrix form, for any composition matrix $\mathbf{x} = [\mathbf{x}_1, \mathbf{x}_2, \cdots, \mathbf{x}_n]'$ composed of n p-dimensional compositions, and $\boldsymbol{\lambda} = (\lambda_1, \lambda_2, \cdots, \lambda_p)' \in \mathbb{R}^p$:

- **Linear combination of $\mathbf{x}_i$ w.r.t. $\boldsymbol{\lambda}$ ($i = 1, 2, \cdots, n$):**

$$\mathbf{x}_i' \odot \boldsymbol{\lambda} = \boldsymbol{\lambda}' \odot \mathbf{x}_i = \lambda_1 \odot \mathbf{x}_{i1} \oplus \lambda_2 \odot \mathbf{x}_{i2} \oplus \cdots \oplus \lambda_p \odot \mathbf{x}_{ip} \in S^D$$

- **Linear combination of $\mathbf{x}$ w.r.t. $\boldsymbol{\lambda}$:**

$$\mathbf{x} \odot \boldsymbol{\lambda} = \left[\left[\mathbf{x}_1' \odot \boldsymbol{\lambda}\right]', \left[\mathbf{x}_2' \odot \boldsymbol{\lambda}\right]', \cdots, \left[\mathbf{x}_n' \odot \boldsymbol{\lambda}\right]' \right]' \in S_n^D$$

- **The ilr transformation of $\mathbf{x}$:**

$$\operatorname{ilr}(\mathbf{x}) = \left(\operatorname{ilr}\left(\mathbf{x}_{(1)}\right), \operatorname{ilr}\left(\mathbf{x}_{(2)}\right), \cdots, \operatorname{ilr}\left(\mathbf{x}_{(p)}\right)\right) \in \mathbb{R}^{(D-1)n_i \times p}$$

$$\mathbf{x}_{(j)} = \left[\mathbf{x}_{1j}', \mathbf{x}_{2j}', \cdots, \mathbf{x}_{nj}'\right]' \in S_n^D \quad (j = 1, 2, \cdots, p)$$

Normal distribution in multiple Simplex space

We say a random vector is **a random composition vector** in S_p^D, if all its values lie in S_p^D.

Definition: *For any random composition vector* $\mathbf{X}$ *in* S_p^D, *we say* $\mathbf{X}$ *satisfies a normal distribution in* S_p^D, *denoted by* $\mathbf{X} \sim \mathcal{N}_S(\boldsymbol{\mu}, \boldsymbol{\Sigma})$, *and* $\boldsymbol{\mu} \in \mathbb{R}^{(D-1)p}$, $\boldsymbol{\Sigma} \in \mathbb{R}^{(D-1)p \times (D-1)p}$, *if the probability density function of* $\mathbf{X}$ *is*

$$f(\mathbf{x}) = \frac{1}{(2\pi)^{(D-1)p/2}|\boldsymbol{\Sigma}|^{1/2}} \exp\Big\{ -\frac{1}{2}\big(\mathrm{ilr}(\mathbf{x}) - \boldsymbol{\mu}\big)' \boldsymbol{\Sigma}^{-1} \big(\mathrm{ilr}(\mathbf{x}) - \boldsymbol{\mu}\big) \Big\},$$

where $\mathbf{x} \in S_p^D$.

Contents

Parameters in CoLMM

In Model (1), pile up all the n_i samples from the i-th individual by row, then for the i-th individual, $i = 1, 2, \cdots, N$, CoLMM can be rewritten as

$$\mathbf{y}_i = \mathbf{x}_i \odot \boldsymbol{\beta} \oplus \mathbf{z}_i' \odot \mathbf{b}_i \oplus \boldsymbol{\varepsilon}_i. \tag{2}$$

Here coefficients of fixed effect $\boldsymbol{\beta}$ are an unknown vector in $\mathbb{R}^p$, coefficients of random effect $\mathbf{b}_i$ are a random vector in $\mathbb{R}^q$, and the error $\boldsymbol{\varepsilon}_i$ is a random composition vector in $S_{n_i}^D$. **We assume** $\mathbf{b}_i$ and $\boldsymbol{\varepsilon}_i$ satisfy normal distributions in $\mathbb{R}^q$ and $S_{n_i}^D$, respectively, i.e.,

$$\begin{aligned} \mathbf{b}_i &\sim \mathcal{N}(\mathbf{0}, \mathbf{G}), \\ \boldsymbol{\varepsilon}_i &\sim \mathcal{N}_S(\mathbf{0}, \sigma^2 \mathbf{I}_{n_i}). \end{aligned}$$

Hence our task is to estimate these parameters, $\boldsymbol{\theta} = (\boldsymbol{\beta}, \mathbf{G}, \sigma^2)$.

Estimate of compositional response

After obtaining the estimates of $\boldsymbol{\beta}$ and $\mathbf{b}_i$, $\widehat{\boldsymbol{\beta}}$ and $\widehat{\mathbf{b}}_i$, the estimates of compositional response, $\widehat{\mathbf{y}}_{ij}$ can be expressed as

- **General form:**

$$\widehat{\mathbf{y}}_{ij} = \mathbf{x}'_{ij} \odot \widehat{\boldsymbol{\beta}} \oplus \mathbf{z}'_{ij} \odot \widehat{\mathbf{b}}_i$$

- **Degenerate form:**

$$\widehat{\mathbf{y}}_{ij} = \mathbf{x}'_{ij} \odot \widehat{\boldsymbol{\beta}}$$

Degenerate form: linear regression for CoDa (CoLM)

When there is no random effect, i.e., $q = 0$, CoLMM is degenerate to linear regression for CoDa (Wang et al., 2013). In degenerate form, the maximum likelihood (ML) estimation of the parameters have explicit solutions.

- **Coefficients of fixed effect:**

$$\widehat{\boldsymbol{\beta}}_{LM} = \Big(\sum_{i=1}^{N} \mathrm{ilr}(\mathbf{x}_i)' \cdot \mathrm{ilr}(\mathbf{x}_i) \Big)^{-1} \Big(\sum_{i=1}^{N} \mathrm{ilr}(\mathbf{x}_i)' \cdot \mathrm{ilr}(\mathbf{y}_i) \Big)$$

- **Parameters of error:**

$$\hat{\sigma}^2_{LM} = \frac{1}{K} \sum_{i=1}^{N} \Big(\mathrm{ilr}(\mathbf{y}_i) - \mathrm{ilr}(\mathbf{x}_i) \cdot \widehat{\boldsymbol{\beta}}_{LM} \Big)' \Big(\mathrm{ilr}(\mathbf{y}_i) - \mathrm{ilr}(\mathbf{x}_i) \cdot \widehat{\boldsymbol{\beta}}_{LM} \Big)$$

$$K = (D-1) \sum_{i=1}^{N} n_i$$

General form: EM algorithm

In general form, we propose an estimate method for parameters in CoLMM within the framework of EM algorithm presented by Laird & Ware (1982).

We consider the all-sample log likelihood function of both compositional response and coefficients of random effect $\log f(\mathbf{y}_i, \mathbf{b}_i|\theta)$, **which gets together with** $S_{n_i}^D$ **and** $\mathbb{R}^q$.

- **E step:**

$$\begin{aligned}\mathbb{E}_{\mathbf{b}_i}\left[\log f(\mathbf{y}_i, \mathbf{b}_i|\theta)\middle|\mathbf{y}_i, \theta\right] &= \mathbb{E}_{\mathbf{b}_i}\left[\log\left[f(\mathbf{y}_i|\mathbf{b}_i, \theta)\cdot\phi(\mathbf{b}_i|\theta)\right]\middle|\mathbf{y}_i, \theta\right] \\ &= \mathbb{E}_{\mathbf{b}_i}\left[\log\left[f(\mathbf{y}_i|\mathbf{b}_i, \theta)\cdot\phi(\mathbf{b}_i|\mathbf{y}_i, \theta)\right]\middle|\theta\right]\end{aligned}$$

- **M step:**

$$\widehat{\theta} = \arg\max_{\theta}\sum_{i=1}^{N}\mathbb{E}_{\mathbf{b}_i}\left[\log f(\mathbf{y}_i, \mathbf{b}_i|\theta)\middle|\mathbf{y}_i, \theta\right]$$

General form: EM algorithm

Lemma 1: *In Model* (2)*, conditioning on* $\boldsymbol{\theta}$*,* $\mathbf{y}_i$ *is normally distributed in* $S_{n_i}^D$*, i.e.,*

$$\mathbf{y}_i \big| \boldsymbol{\theta} \sim \mathcal{N}_S\big(\mathrm{ilr}(\mathbf{x}_i) \cdot \boldsymbol{\beta}, \mathrm{ilr}(\mathbf{z}_i) \cdot \mathbf{G} \cdot \mathrm{ilr}(\mathbf{z}_i)' + \sigma^2 \mathbf{I}_{i*}\big).$$

Lemma 2: *In Model* (2)*, conditioning on* $\mathbf{b}_i$ *and* $\boldsymbol{\theta}$*,* $\mathbf{y}_i$ *is normally distributed in* $S_{n_i}^D$*, i.e.,*

$$\mathbf{y}_i | (\mathbf{b}_i, \boldsymbol{\theta}) \sim \mathcal{N}_S\big(\mathrm{ilr}(\mathbf{x}_i) \cdot \boldsymbol{\beta} + \mathrm{ilr}(\mathbf{z}_i) \cdot \mathbf{b}_i, \sigma^2 \mathbf{I}_{i*}\big).$$

Lemma 3: *In Model* (2)*, conditioning on* $\mathbf{y}_i$ *and* $\boldsymbol{\theta}$*,* $\mathbf{b}_i$ *is normally distributed in* $\mathbb{R}^q$*, i.e.,*

$$\mathbf{b}_i | (\mathbf{y}_i, \boldsymbol{\theta}) = \mathbf{b}_i | (\mathrm{ilr}(\mathbf{y}_i), \boldsymbol{\theta}) \sim \mathcal{N}\big(\boldsymbol{\mu}_i, \boldsymbol{\Phi}_i\big).$$

Here $\mathbf{I}_{i*}$ denotes the unit matrix with dimension $(D-1)n_i$, and

$$\begin{aligned} \boldsymbol{\mu}_i &= \mathbf{G} \cdot \mathrm{ilr}\big(\mathbf{z}_i\big)' \cdot \big(\mathrm{ilr}(\mathbf{z}_i) \cdot \mathbf{G} \cdot \mathrm{ilr}(\mathbf{z}_i)' + \sigma^2 \mathbf{I}_{i*}\big)^{-1} \big(\mathrm{ilr}(\mathbf{y}_i) - \mathrm{ilr}(\mathbf{x}_i) \cdot \boldsymbol{\beta}\big) \\ \boldsymbol{\Phi}_i &= \mathbf{G} - \mathbf{G} \cdot \mathrm{ilr}(\mathbf{z}_i)' \cdot \big(\mathrm{ilr}(\mathbf{z}_i) \cdot \mathbf{G} \cdot \mathrm{ilr}(\mathbf{z}_i)' + \sigma^2 \mathbf{I}_{i*}\big)^{-1} \cdot \mathrm{ilr}(\mathbf{z}_i) \cdot \mathbf{G} \end{aligned}$$

General form: EM algorithm

Let superscript t denote the time of iteration, and $t = 0$ the initial values of iteration.

$$\widehat{\mathbf{\Sigma}}_i^{(t)} = \mathrm{ilr}(\mathbf{z}_i) \cdot \widehat{\mathbf{G}}^{(t)} \cdot \mathrm{ilr}(\mathbf{z}_i)' + \hat{\sigma}^{(t)^2} \mathbf{I}_{i*}$$

$$\widehat{\boldsymbol{\beta}}^{(t)} = \Big(\sum_{i=1}^{N} \mathrm{ilr}(\mathbf{x}_i)' \cdot (\widehat{\mathbf{\Sigma}}_i^{(t)})^{-1} \cdot \mathrm{ilr}(\mathbf{x}_i) \Big)^{-1} \Big(\sum_{i=1}^{N} \mathrm{ilr}(\mathbf{x}_i)' \cdot (\widehat{\mathbf{\Sigma}}_i^{(t)})^{-1} \cdot \mathrm{ilr}(\mathbf{y}_i) \Big)$$

$$\widehat{\mathbf{r}}_i^{(t)} = \mathrm{ilr}(\mathbf{y}_i) - \mathrm{ilr}(\mathbf{x}_i) \cdot \widehat{\boldsymbol{\beta}}^{(t)}$$

$$\widehat{\mathbf{b}}_i^{(t)} = \widehat{\mathbf{G}}^{(t)} \cdot \mathrm{ilr}(\mathbf{z}_i)' \cdot (\widehat{\mathbf{\Sigma}}_i^{(t)})^{-1} \widehat{\mathbf{r}}_i^{(t)}$$

$$\widehat{\mathbf{G}}^{(t+1)} = \frac{1}{N} \sum_{i=1}^{N} \Big(\widehat{\mathbf{b}}_i^{(t)} \widehat{\mathbf{b}}_i^{(t)'} + \widehat{\mathbf{G}}^{(t)} \big(\mathbf{I}_q - \mathrm{ilr}(\mathbf{z}_i)' \cdot (\widehat{\mathbf{\Sigma}}_i^{(t)})^{-1} \cdot \mathrm{ilr}(\mathbf{z}_i) \cdot \widehat{\mathbf{G}}^{(t)} \big) \Big)$$

$$\hat{\sigma}^{(t+1)^2} = \frac{1}{K} \sum_{i=1}^{N} \Big((\widehat{\mathbf{r}}_i^{(t)} - \mathrm{ilr}(\mathbf{z}_i) \cdot \widehat{\mathbf{b}}_i^{(t)})' (\widehat{\mathbf{r}}_i^{(t)} - \mathrm{ilr}(\mathbf{z}_i) \cdot \widehat{\mathbf{b}}_i^{(t)}) + \hat{\sigma}^{(t)^2} \mathrm{tr}\big(\mathbf{I}_{i*} - \hat{\sigma}^{(t)^2} (\widehat{\mathbf{\Sigma}}_i^{(t)})^{-1} \big) \Big)$$

$$t = t + 1$$

Repeat the aforementioned process until convergence or beyond the iteration limit.

Setting of initial values

Suggested by Laird et al. (1987), we set the initial values as

$$\begin{aligned}
\widehat{\boldsymbol{\beta}}^{(0)} &= \widehat{\boldsymbol{\beta}}_{LM} \\
\hat{\sigma}^{(0)^2} &= \frac{1}{M}\sum_{i=1}^{N}\left(\mathrm{ilr}(\mathbf{y}_i) - \mathrm{ilr}(\mathbf{z}_i)\cdot\widehat{\mathbf{b}}_i^{(0)}\right)'\widehat{\mathbf{r}}_i^{(0)} \\
\widehat{\mathbf{G}}^{(0)} &= \frac{1}{N}\sum_{i=1}^{N}\left(\widehat{\mathbf{b}}_i^{(0)}\widehat{\mathbf{b}}_i^{(0)'} - \hat{\sigma}^{(0)^2}\left(\mathrm{ilr}(\mathbf{z}_i)'\cdot\mathrm{ilr}(\mathbf{z}_i)\right)^{-1}\right)
\end{aligned}$$

Here $M = K - (N-1)q - p$ and

$$\widehat{\mathbf{b}}_i^{(0)} = \left(\mathrm{ilr}(\mathbf{z}_i)'\cdot\mathrm{ilr}(\mathbf{z}_i)\right)^{-1}\cdot\mathrm{ilr}(\mathbf{z}_i)'\cdot\widehat{\mathbf{r}}_i^{(0)}.$$

Independency

In the former process, we performed the estimation by the ilr coordinates. **Here we need to show that** the proposed estimate method does not depend on the choice of the ilr transformation, or equivalently the contact matrix $\mathbf{\Phi}$.

Recall that the ilr transformation and the corresponding contact matrix $\mathbf{\Phi}$ satisfy

$$\begin{aligned} \mathrm{ilr}(\mathbf{x}) &= \mathbf{\Phi} \cdot \log(\mathbf{x}) \\ \mathbf{\Phi}'\mathbf{\Phi} &= \mathbf{I}_{D-1} \end{aligned}$$

Hence we show that both "ilr" and $\mathbf{\Phi}$ in the algorithm can be substituted by just "log".

Independency

- **Initial values**

$$
\begin{aligned}
\widehat{\boldsymbol{\beta}}^{(0)} &= \Big(\sum_{i=1}^{N} \log(\mathbf{x}_i)' \boldsymbol{\Phi}' \boldsymbol{\Phi} \log(\mathbf{x}_i)\Big)^{-1} \Big(\sum_{i=1}^{N} \log(\mathbf{x}_i)' \boldsymbol{\Phi}' \boldsymbol{\Phi} \log(\mathbf{y}_i)\Big) \\
&= \Big(\sum_{i=1}^{N} \log(\mathbf{x}_i)' \log(\mathbf{x}_i)\Big)^{-1} \Big(\sum_{i=1}^{N} \log(\mathbf{x}_i)' \log(\mathbf{y}_i)\Big) \\
\hat{\sigma}^{(0)^2} &= \frac{1}{M} \sum_{i=1}^{N} \Big(\log(\mathbf{y}_i) - \log(\mathbf{x}_i)\widehat{\mathbf{b}}_i^{(0)}\Big)' \Big(\log(\mathbf{y}_i) - \log(\mathbf{x}_i)\widehat{\boldsymbol{\beta}}_i^{(0)}\Big) \\
\widehat{\mathbf{G}}^{(0)} &= \frac{1}{N} \sum_{i=1}^{N} \Big(\widehat{\mathbf{b}}_i^{(0)} \widehat{\mathbf{b}}_i^{(0)'} - \hat{\sigma}^{(0)^2} \big(\log(\mathbf{z}_i)' \log(\mathbf{z}_i)\big)^{-1}\Big) \\
\widehat{\mathbf{b}}_i^{(0)} &= \big(\log(\mathbf{z}_i)' \log(\mathbf{z}_i)\big)^{-1} \log(\mathbf{z}_i)' \Big(\log(\mathbf{y}_i) - \log(\mathbf{x}_i)\widehat{\boldsymbol{\beta}}_i^{(0)}\Big)
\end{aligned}
$$

Independency

- **Iteration**

Notice that

$$(\widehat{\boldsymbol{\Sigma}}_i^{(t)})^{-1} = \frac{1}{\sigma^2}\mathbf{I}_{i*} - \frac{1}{\sigma^4}\boldsymbol{\Phi}\log(\mathbf{z}_i)\big(\widehat{\mathbf{G}}^{-1} + \frac{1}{\sigma^2}\log(\mathbf{z}_i)'\log(\mathbf{z}_i)\big)^{-1}\log(\mathbf{z}_i)'\boldsymbol{\Phi}'$$

and

$$\mathrm{tr}\big(\mathbf{I}_{i*} - \hat{\sigma}^{(t)^2}(\widehat{\boldsymbol{\Sigma}}_i^{(t)})^{-1}\big) = \mathrm{tr}\Big(\frac{1}{\sigma^2}\mathbf{I}_{i*} - \frac{1}{\sigma^4}\log(\mathbf{z}_i)\big(\widehat{\mathbf{G}}^{-1} + \frac{1}{\sigma^2}\log(\mathbf{z}_i)'\log(\mathbf{z}_i)\big)^{-1}\log(\mathbf{z}_i)'\Big)$$

Hence the iteration does not depend on $\boldsymbol{\Phi}$. Both the initial values and the iteration show that the estimate results $\widehat{\boldsymbol{\theta}} = (\widehat{\boldsymbol{\beta}}, \widehat{\mathbf{G}}, \hat{\sigma}^2)$ are unique.

Consistency

It is obvious that the proposed estimate method is consistent with the existing estimate in CoLM when $q = 0$.

- **Coefficients of random effect:**

$$\widehat{\boldsymbol{\Sigma}}_i^{(t)} = \hat{\sigma}^{(t)^2} \mathbf{I}_{i*} \quad \Rightarrow \quad \widehat{\boldsymbol{\beta}}^{(t)} \equiv \widehat{\boldsymbol{\beta}}_{LM}$$

- **Parameter of error:**

$$\hat{\sigma}^{(t+1)^2} = \frac{1}{K} \sum_{i=1}^{N} \left(\mathrm{ilr}(\mathbf{y}_i) - \mathrm{ilr}(\mathbf{x}_i) \cdot \widehat{\boldsymbol{\beta}}^{(t)}\right)' \left(\mathrm{ilr}(\mathbf{y}_i) - \mathrm{ilr}(\mathbf{x}_i) \cdot \widehat{\boldsymbol{\beta}}^{(t)}\right) \equiv \hat{\sigma}^2_{LM}$$

Contents

Generation

We generated data from Model (2). $\mathbf{y}_i = \mathbf{x}_i \odot \boldsymbol{\beta} \oplus \mathbf{z}_i \odot \mathbf{b}_i \oplus \boldsymbol{\varepsilon}_i$

- $D = 4 \quad p = 5 \quad q = 3 \quad \boldsymbol{\beta} = (2, 1, -1, -2, 0)'$
- $\mathbf{x}_i \sim \mathcal{NS}(\mathbf{0}, \mathbf{I}_{15}) \quad \mathbf{z}_i = \left[\mathbf{x}'_{i1}, \mathbf{x}'_{i2}, \mathbf{x}'_{i3}\right]'$
- $\mathbf{x}_i \perp \boldsymbol{\varepsilon}_i \sim \mathcal{NS}(\mathbf{0}, \sigma^2 \mathbf{I}_3) \quad \sigma^2 \in \{0.5, 1, 1.5\}$
- $\mathbf{b}_i \sim \mathcal{N}(\mathbf{0}, \mathbf{G})$

$$\mathbf{G} = \begin{pmatrix} g_{11} & g_{12} & g_{13} \\ g_{21} & g_{22} & g_{23} \\ g_{31} & g_{32} & g_{33} \end{pmatrix} = \begin{pmatrix} 9 & 4.8 & 0.6 \\ 4.8 & 4 & 1 \\ 0.6 & 1 & 1 \end{pmatrix}$$

- $(N, n_i) \in \{(30, 5), (60, 5), (60, 10)\}$
- Repeated 500 times and conducted CoLMM and CoLM, respectively.

Evaluation indicators

- **Mean absolute percentage error for CoDa (CoMAPE):**

$$\mathrm{CoMAPE} = \frac{D-1}{K} \sum_{i=1}^{N} \sum_{j=1}^{n_i} \frac{||\mathbf{y}_{ij} \ominus \widehat{\mathbf{y}}_{ij}||_S^2}{||\mathbf{y}_{ij}||_S^2}$$

CoMAPE is nonnegative. **A lower CoMAPE indicates a better model.**

- **Goodness of fit for CoDa (χ^2):**

$$\chi^2 = \frac{\sum_{i=1}^{N} \sum_{j=1}^{n_i} ||\widehat{\mathbf{y}}_{ij}||_S^2}{\sum_{i=1}^{N} \sum_{j=1}^{n_i} ||\mathbf{y}_{ij}||_S^2}$$

χ^2 values in $[0, 1]$. χ^2 **that is closer to 1 also indicates a better model.**

Table: Means and standard errors (in brackets) of estimates of both coefficients of fixed effect and parameters of error, and CoMAPE and χ^2 of the compositional response.

(N, n_i)	**Model**	$\hat{\beta}_1$	$\hat{\beta}_2$	$\hat{\beta}_3$	$\hat{\beta}_4$	$\hat{\beta}_5$	**CMAPE**	χ^2	$\hat{\sigma}^2$
Oracle		2	1	-1	-2	0	0	1	0
					$\sigma = 0.5$				
(30, 5)	CoLM	1.961 (0.635)	0.991 (0.433)	−1.007 (0.267)	−1.999 (0.177)	0.006 (0.169)	1.355 (1.137)	0.434 (0.092)	3.664 (0.444)
	CoLMM	1.977 (0.565)	0.99 (0.373)	−0.998 (0.183)	−1.999 (0.027)	0 (0.026)	0.046 (0.071)	0.991 (0.002)	0.499 (0.018)
(60, 5)	CoLM	1.998 (0.403)	1.004 (0.322)	−0.99 (0.187)	−2.001 (0.126)	0.014 (0.123)	1.326 (0.638)	0.426 (0.06)	3.709 (0.302)
	CoLMM	1.989 (0.372)	1 (0.273)	−0.999 (0.134)	−2 (0.019)	0.001 (0.019)	0.041 (0.013)	0.992 (0.001)	0.499 (0.013)
(60, 10)	CoLM	1.966 (0.424)	0.988 (0.297)	−1 (0.166)	−1.998 (0.093)	0.002 (0.091)	1.366 (0.623)	0.42 (0.061)	3.718 (0.295)
	CoLMM	1.962 (0.409)	0.984 (0.279)	−0.995 (0.132)	−2 (0.013)	0 (0.013)	0.049 (0.017)	0.99 (0.001)	0.499 (0.009)
					$\sigma = 1.5$				
(30, 5)	CoLM	1.964 (0.64)	0.989 (0.438)	−1.008 (0.276)	−1.997 (0.19)	0.006 (0.183)	1.322 (0.647)	0.401 (0.087)	3.925 (0.421)
	CoLMM	1.981 (0.571)	0.987 (0.379)	−0.998 (0.203)	−1.998 (0.079)	0.001 (0.08)	0.284 (0.12)	0.925 (0.025)	1.5 (0.114)
(60, 5)	CoLM	1.998 (0.407)	1.005 (0.325)	−0.993 (0.196)	−2.001 (0.134)	0.015 (0.131)	1.346 (0.439)	0.394 (0.057)	3.97 (0.287)
	CoLMM	1.99 (0.377)	1.001 (0.278)	−1.001 (0.147)	−2.001 (0.055)	0.003 (0.055)	0.298 (0.146)	0.928 (0.01)	1.497 (0.04)
(60, 10)	CoLM	1.967 (0.425)	0.988 (0.3)	−0.999 (0.169)	−1.998 (0.101)	0.002 (0.097)	1.329 (0.352)	0.388 (0.057)	3.978 (0.277)
	CoLMM	1.963 (0.411)	0.985 (0.282)	−0.995 (0.136)	−2 (0.038)	0 (0.038)	0.317 (0.091)	0.921 (0.01)	1.497 (0.028)

Table: Means and standard errors (in brackets) of the bias in the covariance matrix of the coefficients of random effect. Only CoLMMs were considered here, $r_{ij} = g_{ij} - \hat{g}_{ij}$ denotes the bias of the element in the i-th row and the j-th column.

(N, n_i)	r_{11}	r_{21}	r_{22}	r_{31}	r_{32}	r_{33}
			$\sigma = 0.5$			
$(30, 5)$	0.31 (0.185)	0.159 (0.102)	0.019 (0.086)	0.132 (0.113)	0.033 (0.062)	0.036 (0.056)
$(60, 5)$	0.147 (0.123)	0.08 (0.071)	0.01 (0.062)	0.069 (0.08)	0.017 (0.043)	0.016 (0.038)
$(60, 10)$	0.149 (0.078)	0.077 (0.048)	0.009 (0.038)	0.065 (0.05)	0.017 (0.027)	0.019 (0.025)
			$\sigma = 1$			
$(30, 5)$	0.318 (0.337)	0.159 (0.19)	0.019 (0.168)	0.134 (0.217)	0.033 (0.122)	0.04 (0.111)
$(60, 5)$	0.145 (0.244)	0.082 (0.139)	0.01 (0.121)	0.072 (0.157)	0.017 (0.086)	0.016 (0.076)
$(60, 10)$	0.148 (0.147)	0.075 (0.092)	0.008 (0.075)	0.064 (0.098)	0.017 (0.054)	0.021 (0.05)
			$\sigma = 1.5$			
$(30, 5)$	0.348 (0.686)	0.17 (0.388)	0.02 (0.249)	0.144 (0.347)	0.034 (0.182)	0.047 (0.169)
$(60, 5)$	0.145 (0.363)	0.083 (0.207)	0.01 (0.18)	0.234 (0.075)	0.018 (0.13)	0.017 (0.116)
$(60, 10)$	0.147 (0.218)	0.072 (0.137)	0.008 (0.133)	0.063 (0.147)	0.017 (0.08)	0.024 (0.075)

Simulation results

From two tables, the results can be summarized that

- Both CoLMM and CoLM can estimate the coefficients of fixed effect well, with CoLMM's more stable.
- CoLM performed bad in fitting the compositional response and almost failed.
- CoLMM performed well in fitting the compositional response and captured the differences of individuals.

In general, our proposed CoLMM can deal with the case that individuals are different in the population, estimate the parameters well and fit the compositional response effectively.

Contents

Data process

From the integrate perspective of **three industries**, we consider the correlation of some structural economic indicators in China. Original data from 2010 to 2015 can be obtained from *China Statistical Yearbook*.

- Gross domestic product (GDP, $\mathbf{y}$)
- Total investment in fixed assets (INVT, $\mathbf{x}_1$)
- Urban unit employment (UUE, $\mathbf{x}_2$)
- Wages of urban unit employment (WUUE, $\mathbf{x}_3$)
- Data table:

$$\text{“31 Areas} \times \text{6 Years} \times \text{3 Indicators”}$$

Modeling results

Table: Fitting and out-of-sample performance of CoLM and CoLMM with different settings of random effect. "∅" denotes CoLM. "INVT", "UUE" and "WUUE" denote the related coefficients of fixed effect, respectively. "CV" denotes the means and standard errors (in brackets) of the CoMAPE of the cross-validation method, where the leave-one-off method was introduced. "⋆" indicates the best CoLMM.

Random effect	INVT	UUE	WUUE	CMAPE	χ^2	CV
$\emptyset$	0.366	0.364	-0.157	0.235	0.854	0.242 (0.486)
$(\mathbf{x}_1)$	0.11	0.168	0.169	0.055	0.969	0.058 (0.07)
$(\mathbf{x}_2)$	0.189	0.035	0.231	0.066	0.971	0.069 (0.096)
$(\mathbf{x}_3)$	0.196	0.119	0.158	0.06	0.972	0.063 (0.076)
$(\mathbf{x}_1, \mathbf{x}_2)$	0.244	0.171	0.057	0.035	0.983	0.041 (0.056)
$(\mathbf{x}_1, \mathbf{x}_3)$	0.243	0.15	0.079	0.036	0.984	0.042 (0.06)
$(\mathbf{x}_2, \mathbf{x}_3)$	0.24	0.432	-0.187	0.034	0.983	0.041 (0.051)
$(\mathbf{x}_1, \mathbf{x}_2, \mathbf{x}_3)^\star$	0.268	0.325	-0.104	0.022	0.99	0.031 (0.044)

About the best model

According to the cross-validation result, we chose the CoLMM with random effect having all the three indicators as the final model.

Table: Elements of the covariance matrix of coefficients of random effect in the best CoLMM.

	INVT	UUE	WUUE
INVT	0.209	0.422	-0.492
UUE		1.675	-1.705
WUUE			1.791

In China, the difference of INVT's influence on GDP is relatively small, while UUE's and WUUE's are relatively large, the variances of which reach 1.675 and 1.791, respectively. For instance, the final models in Beijing and Shanghai are

$$\text{Beijing} \qquad \widehat{\mathbf{y}} = 0.049 \odot \mathbf{x}_1 \ominus 0.037 \odot \mathbf{x}_2 \oplus 0.76 \odot \mathbf{x}_3$$

$$\text{Shanghai} \qquad \widehat{\mathbf{y}} = 0.384 \odot \mathbf{x}_1 \oplus 0.777 \odot \mathbf{x}_2 \ominus 0.282 \odot \mathbf{x}_3$$

Contents

Further work

Now We have proposed the ML estimation for CoLMM by EM algorithm, and have shown some properties of the estimation procedure. However something remains to be finished.

- Restricted maximum likelihood (REML) estimation
- Statistical inference for the estimates of parameters

Here we consider compositional data. **What about other types of symbolic data?**

- Interval data: centers and ranges are dependent on individuals
- Histogram: part of quantiles are dependent on individuals

THANKS!