Calibration between log-ratios of parts of compositional data using linear models

S. Donevska E. Fišerová K. Hron

Palacký University in Olomouc, Czech Republic

ROBUST 2012

Compositional data analysis

Compositional data (CoDa) = quantitative descriptions of parts of some whole, thus as data carring only **relative information**.
Simplex with the Aitchison geometry= the sample space of CoDa,

$$\mathcal{S}^D = \{\mathbf{x} = (x_1, \ldots, x_D)',\ x_i > 0,\ \sum_{i=1}^{D} x_i = \kappa\}.$$

$\Rightarrow$ Aitchison geometry forms a **vector space structure of the simplex**.

ILR transformation= isometric mapping from $\mathcal{S}^D$ to $\mathbb{R}^{D-1}$.

$\Rightarrow$ **Adventage**: Using ilr transformation we obtain **orthonormal coordinates** on the $\mathbb{R}^{D-1}$,

$$\mathrm{ilr}(\mathbf{x}) = \mathbf{z} = (z_1, \ldots, z_{D-1})',\ z_i = \sqrt{\frac{i}{i+1}} \ln \frac{\sqrt[i]{\prod_{j=1}^{i} x_j}}{x_{i+1}}.$$

Compositional data analysis

Compositional variation array=tool for exploratory compositional data analysis,

$$\mathbf{V} = \begin{pmatrix} 0 & \operatorname{var}\ln\left(\frac{x_1}{x_2}\right) & \operatorname{var}\ln\left(\frac{x_1}{x_3}\right) & \cdots & \operatorname{var}\ln\left(\frac{x_1}{x_D}\right) \\ \mathrm{E}\ln\left(\frac{x_2}{x_1}\right) & 0 & \operatorname{var}\ln\left(\frac{x_2}{x_3}\right) & \cdots & \operatorname{var}\ln\left(\frac{x_2}{x_D}\right) \\ \vdots & \vdots & \vdots & \vdots & \vdots \\ \mathrm{E}\ln\left(\frac{x_D}{x_1}\right) & \mathrm{E}\ln\left(\frac{x_D}{x_2}\right) & \mathrm{E}\ln\left(\frac{x_D}{x_3}\right) & \cdots & 0 \end{pmatrix}.$$

Properties:

⇒ **Log-ratio variances** satisfy the **symmetric property**, i.e.,

$$\operatorname{var}\ln\left(\frac{x_i}{x_j}\right) = \operatorname{var}\left(-\ln\left(\frac{x_j}{x_i}\right)\right).$$

⇒ For the **log-ratio means** the **triangular equality** holds, i.e.,

$$\mathrm{E}\ln\left(\frac{x_j}{x_k}\right) = \mathrm{E}\ln\left(\frac{x_j}{x_i}\right) + \mathrm{E}\ln\left(\frac{x_i}{x_k}\right).$$

Calibration problem for CoDa

Task: For D-part composition we split the calibration problem into $\frac{D(D-1)}{2}$ partial calibration problems.

$\Rightarrow$ **This means that we will calibrate each of the 2-part subcompositions of the given composition.**

Consideration: We have n different objects that have D properties which are measured on two different measuring devices A and B, that measure with the same imprecision.

Data matrices: Ilr transformed 2-part subcompositions $(x_r, x_s)'$ resp. $(y_r, y_s)'$ corresponding to the measurement results from A resp. B, multiplied by $\sqrt{2}$ create the data matrices,

$$(\mathbf{Z}_k^A, \mathbf{Z}_k^B)^{(r,s)} = \begin{pmatrix} \ln \frac{x_{1r}}{x_{1s}} & \ln \frac{y_{1r}}{y_{1s}} \\ \ln \frac{x_{2r}}{x_{2s}} & \ln \frac{y_{nr}}{y_{ns}} \\ \vdots & \vdots \\ \ln \frac{x_{nr}}{x_{ns}} & \ln \frac{y_{nr}}{y_{ns}} \end{pmatrix}, k = 1, \ldots, \frac{D(D-1)}{2}.$$

Calibration problem for CoDa

Linear model with type-II constraints:

$$
\begin{aligned}
\begin{pmatrix} \mathbf{z}_k^A \\ \mathbf{z}_k^B \end{pmatrix} &= \begin{pmatrix} \boldsymbol{\mu}_k \\ \boldsymbol{\nu}_k \end{pmatrix} + \varepsilon, \quad \operatorname{var}(\varepsilon) = \sigma^2 \mathbf{I}, & (1)\\
\boldsymbol{\nu}_k &= \beta_{1k}\mathbf{1}_n + \beta_{2k}\boldsymbol{\mu}_k, & (2)
\end{aligned}
$$

$$k = 1, \ldots, \frac{D(D-1)}{2}$$

- $\mathbf{z}_k^i$, $i = A, B$ is n-dimensional random vector created by realization of the data $\mathbf{Z}_k^i$, $i = A, B$,
- $\boldsymbol{\mu}_k = (\mu_{1k}, \ldots, \mu_{nk})'$, $\boldsymbol{\nu}_k = (\nu_{1k}, \ldots, \nu_{nk})'$ are an errorless recordings of $\mathbf{z}_k^A$ and $\mathbf{z}_k^B$ resp.,
- $\boldsymbol{\nu}_k = \beta_{1k}\mathbf{1}_n + \beta_{2k}\boldsymbol{\mu}_k$, is the calibration line,
  - $\rightarrow$ $\boldsymbol{\mu}_k$ and $\boldsymbol{\nu}_k$ are realized independently with an error $\sigma > 0$.
- β_{1k} and β_{2k} are unknown coefficients that specify the intercept and the slope of the calibration line.

$\Rightarrow$ $\boldsymbol{\mu}_k, \boldsymbol{\nu}_k, \beta_{1k}$ and β_{2k} need to be estimated in an **iterative manner**.

$\Rightarrow$ $\widehat{\beta}_{1k}$ and $\widehat{\beta}_{2k}$ converge to the orthogonal least squares estimates.

Calibration problem for CoDa

$\Rightarrow$ The unbiased estimator of the **unknown variance** σ^2 is

$$\widehat{\sigma}^2 = \frac{\left(\mathbf{z}_k^A - \widehat{\boldsymbol{\mu}}_k\right)' \left(\mathbf{z}_k^A - \widehat{\boldsymbol{\mu}}_k\right) + \left(\mathbf{z}_k^B - \widehat{\boldsymbol{\nu}}_k\right)' \left(\mathbf{z}_k^B - \widehat{\boldsymbol{\nu}}_k\right)}{n-2}. \tag{3}$$

Matrices of the predicted averages $\mathbf{M}^{(j)}$, $j = 1, 2$:

$$\mathbf{M}^{(1)} = \begin{pmatrix} 0 & \overline{\ln \frac{\mathbf{x}_1}{\mathbf{x}_2}} & \overline{\ln \frac{\mathbf{x}_1}{\mathbf{x}_3}} & \cdots & \overline{\ln \frac{\mathbf{x}_1}{\mathbf{x}_D}} \\ \overline{\ln \frac{\mathbf{x}_2}{\mathbf{x}_1}} & 0 & \overline{\ln \frac{\mathbf{x}_2}{\mathbf{x}_3}} & \cdots & \overline{\ln \frac{\mathbf{x}_2}{\mathbf{x}_D}} \\ \vdots & \vdots & \vdots & \vdots & \vdots \\ \overline{\ln \frac{\mathbf{x}_D}{\mathbf{x}_1}} & \overline{\ln \frac{\mathbf{x}_D}{\mathbf{x}_2}} & \overline{\ln \frac{\mathbf{x}_D}{\mathbf{x}_3}} & \cdots & 0 \end{pmatrix},$$

- $\overline{\ln \frac{\mathbf{x}_r}{\mathbf{x}_s}}$, $r, s = 1, \ldots, D$ is the predicted average for the model (1)-(2), i.e.,

$$\overline{\ln \frac{\mathbf{x}_r}{\mathbf{x}_s}} = \widehat{\beta}_1^{rs(1)} + \widehat{\beta}_2^{rs(1)} \frac{1}{n} \sum_{i=1}^{n} \ln \frac{x_{ir}}{x_{is}},$$

Calibration problem for CoDa

$$\mathbf{M}^{(2)} = \begin{pmatrix} 0 & \overline{\ln \frac{\mathbf{y}_1}{\mathbf{y}_2}} & \overline{\ln \frac{\mathbf{y}_1}{\mathbf{y}_3}} & \cdots & \overline{\ln \frac{\mathbf{y}_1}{\mathbf{y}_D}} \\ \overline{\ln \frac{\mathbf{y}_2}{\mathbf{y}_1}} & 0 & \overline{\ln \frac{\mathbf{y}_2}{\mathbf{y}_3}} & \cdots & \overline{\ln \frac{\mathbf{y}_2}{\mathbf{y}_D}} \\ \vdots & \vdots & \vdots & \vdots & \vdots \\ \overline{\ln \frac{\mathbf{y}_D}{\mathbf{y}_1}} & \overline{\ln \frac{\mathbf{y}_D}{\mathbf{y}_2}} & \overline{\ln \frac{\mathbf{y}_D}{\mathbf{y}_3}} & \cdots & 0 \end{pmatrix},$$

- $\overline{\ln \frac{\mathbf{y}_r}{\mathbf{y}_s}}$, $r, s = 1, \ldots, D$ is the predicted average for the linear model with type II constraint

$$\begin{aligned} \begin{pmatrix} \mathbf{z}_k^B \\ \mathbf{z}_k^A \end{pmatrix} &= \begin{pmatrix} \boldsymbol{\nu}_k \\ \boldsymbol{\mu}_k \end{pmatrix} + \varepsilon, \quad \mathrm{var}(\varepsilon) = \sigma^2 \mathbf{I}, \\ \boldsymbol{\mu}_k &= \beta_{1k} \mathbf{1}_n + \beta_{2k} \boldsymbol{\nu}_k, \end{aligned}$$

i.e., $\overline{\ln \frac{\mathbf{y}_r}{\mathbf{y}_s}} = \widehat{\beta}_1^{rs(2)} + \widehat{\beta}_2^{rs(2)} \frac{1}{n} \sum_{i=1}^{n} \ln \frac{y_{ir}}{y_{is}}$,

$\Rightarrow$ $\mathbf{M}^{(j)}$, $j = 1, 2$ are **asymmetric matrices** and for their elements the **triangular equality** holds.

Calibration problem for CoDa

Matrix of residual variances T:

$$\mathbf{T} = \begin{pmatrix} 0 & \widehat{\sigma}^2_{12} & \widehat{\sigma}^2_{13} & \cdots & \widehat{\sigma}^2_{1D} \\ \widehat{\sigma}^2_{21} & 0 & \widehat{\sigma}^2_{23} & \cdots & \widehat{\sigma}^2_{2D} \\ \vdots & \vdots & \vdots & \vdots & \vdots \\ \widehat{\sigma}^2_{D1} & \widehat{\sigma}^2_{D2} & \widehat{\sigma}^2_{D3} & \cdots & 0 \end{pmatrix},$$

- $\widehat{\sigma}^2_{rs}$ $r, s = 1, \ldots, D$ is the estimate of the residual variance in the model (1)-(2) corresponding to the log-ratios of the parts $(x_r, x_s)'$, calculated according to (3).

$\Rightarrow$ **T** is **symmetric matrix**.

Ilustrative example

We consider the White Blood Cells data set [1] of 30 samples obtained by two different methods: microscopic inspection and image analysis.

Consists of three parts:

- granulocytes (= part x_1),
- lymphocytes (= part x_2),
- monocytes (= part x_3).

$\Rightarrow$ Calibration lines are estimated by the iterative algorithm described in [2], and they are determined with a high precision.

k	calibration line / standard errors of $(\widehat{\beta}_{1k}, \widehat{\beta}_{2k})$	iterations
1	$\mathbf{z}_2^{(1,2)} = 0.1719 + 1.0232\mathbf{z}_1^{(1,2)}$ (0.0532, 0.0334)	9
2	$\mathbf{z}_2^{(1,3)} = 0.0647 + 0.9972\mathbf{z}_1^{(1,3)}$ (0.0606, 0.0210)	7
3	$\mathbf{z}_2^{(2,3)} = -0.1332 + 0.9971\mathbf{z}_1^{(2,3)}$ (0.0458, 0.0228)	7

Ilustrative example

Testing hypothesis (for given r,s) [3]:

- Both methods measure with the same precision of 0.2, (prescribed precision of devices), i.e.,

$$\mathrm{H}_0\colon \sigma^2_{rs} = 0.2^2 \text{ v.s. } \mathrm{H}_A\colon \sigma^2_{rs} \neq 0.2^2.$$

  - Under H_0: $\widehat{\sigma}^2_{rs} \dfrac{n-2}{\sigma^2_{rs}} \sim \chi^2_{n-2}$.

$\Rightarrow$ In our example, on the significance level 0.05 we accept the H_0, i.e., the both instruments measure with the same precision 0.2.

- The results obtained from the both methods do not differ, i.e.,

$$\mathrm{H}_0\colon \mu_{rs} = \nu_{rs} \text{ v.s. } \mathrm{H}_A\colon \mu_{rs} \neq \nu_{rs}.$$

  - Under H_0: $T = \dfrac{\overline{\ln \frac{\mathbf{x}_r}{\mathbf{x}_r}} - \overline{\ln \frac{\mathbf{y}_r}{\mathbf{y}_s}} - (\mu_{rs} - \nu_{rs})}{\sqrt{(n-1)\mathrm{s}^2_{\ln \frac{\mathbf{x}_r}{\mathbf{x}_s}} + (n-1)\mathrm{s}^2_{\ln \frac{\mathbf{y}_r}{\mathbf{y}_s}}}} \sqrt{n(n-1)} \sim \mathrm{t}_{2(n-1)}$,

$\rightarrow$ $\mathrm{s}^2_{\ln \frac{\mathbf{x}_r}{\mathbf{x}_s}}$ and $\mathrm{s}^2_{\ln \frac{\mathbf{y}_r}{\mathbf{y}_s}}$ are sample variances.

$\Rightarrow$ Again we did not reject the H_0 on the significance level 0.05, which means that the both methods give us the same results.

References

1. Aitchison, J.: The Statistical Analysis of Compositional Data. London: Chapman and Hall, 1986.
2. Fišerová, E., Hron, K.: Total least squares solution for compositional data using linear models. *Journal of Applied Statistics* **37**, 7 (2010), 1137–1152.
3. Fišerová, E., Hron, K.: Statistical inference in orthogonal regression for three-part compositional data using a linear model with type-II constraints.*Communications in Statistics - Theory and Methods*, **41**, (2012), 2367–2385.