CALIBRATION BETWEEN LOG-RATIOS OF PARTS OF COMPOSITIONAL DATA USING LINEAR MODELS

Sandra Donevska, Eva Fišerová and Karel Hron
sdonevska@seznam.cz
Department of Mathematical Analysis and Applications of Mathematics,
Palacký University in Olomouc, Czech Republic

Calibration curve describes the relation between the errorless measurement results obtained by measuring the same object on two different measuring devises (techniques). The problem of fitting a calibration curve is known as a calibration problem in the statistical literature. Here we will consider the measurement data to be compositional. Compositional data are defined as quantitative descriptions of parts of some whole, thus as data carrying only relative information. This kind of multivariate data require different treatment in sense of accomplishing standard statistical techniques, what is the case of the calibration. We will present here the calibration line problem solved by the linear model with type-II constraints for compositional data. Further our aim is to preform statistical inference and also to find analogy between the compositional variation array and the matrices of the predicted values and the matrix of the residual values.

COMPOSITIONAL DATA

- The sample space of D-part compositional data is **the simplex**,

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

  – The closuring to the constant κ, usually chosen as 1 or 100 (for expressing in proportions or percentages), is a consequence of **the closure operation**

$$\mathcal{C}(\mathbf{x}) = \left(\frac{\kappa x_1}{\sum_{i=1}^{D} x_i}, \dots, \frac{\kappa x_D}{\sum_{i=1}^{D} x_i}\right)'.$$

  – **Aitchison geometry** forms the Euclidean structure of the simplex, it is represented by the operations of *perturbation*, *power transformation* and *the Aitchison inner product*

$$\mathbf{x} \oplus \mathbf{y} = \mathcal{C}\left(x_1 y_1, x_2 y_2, ..., x_D y_D\right)', \alpha \odot \mathbf{x} = \mathcal{C}\left(x_1^\alpha, x_2^\alpha, ..., x_D^\alpha\right)',$$

$$\langle \mathbf{x}, \mathbf{y} \rangle_a = \frac{1}{2D} \sum_{i=1}^{D} \sum_{j=1}^{D} \ln \frac{x_i}{x_j} \ln \frac{y_i}{y_j}.$$

- Using **the isometric log-ratio (ilr) transformation** we obtain orthonormal coordinates on the Euclidean real space

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

- Tool for exploratory compositional data analysis is **the compositional variation array** [1]

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

  ⇒ Log-ratio variances satisfy the symmetric property i.e. $\mathrm{var}\ln\left(\frac{x_i}{x_j}\right) = \mathrm{var}\left(-\ln\left(\frac{x_j}{x_i}\right)\right)$.

  ⇒ For the log-ratio means holds the triangular equality 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 THE COMPOSITIONAL DATA

- For D-part compositional data the task is to split the calibration problem into $\frac{D(D-1)}{2}$ partial calibration problems.

  ⇒ This means that we will calibrate each of the 2-part subcompositions of the given compositional data.

- Consider 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.

- Ilr transformed two-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,

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

- Linear model with type-II constraints is given by

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

$$\boldsymbol{\nu}_k = \beta_{1k}\mathbf{1}_n + \beta_{2k}\boldsymbol{\mu}_k, \tag{2}$$

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

  – $\mathbf{z}_k^i$, $i = A, B$ is n-dimensional random vector created by realization of the data $\mathbf{Z}_k^i$, $i = A, B$,

  – $\boldsymbol{\mu}_k = (\mu_{1k}, \dots, \mu_{nk})'$, $\boldsymbol{\nu}_k = (\nu_{1k}, \dots, \nu_{nk})'$ are an errorless recording of $\mathbf{z}_k^A$ and $\mathbf{z}_k^B$ respectively,

  – $\boldsymbol{\nu}_k = \beta_{1k}\mathbf{1}_n + \beta_{2k}\boldsymbol{\mu}_k$, is the calibration line,

  ⇒ $\boldsymbol{\mu}_k$ and $\boldsymbol{\nu}_k$ are independent and realized with an error $\sigma > 0$.

  – β_{1k} and β_{2k} are unknown coefficients that specify the intercept and the slope of the calibration line.

- Formulas for BLUE of $\boldsymbol{\mu}$, $\boldsymbol{\nu}$, β_1 and β_2 in the linearized model are given in [2], [3], they need to be estimated in an iterative manner.

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

- 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 $\mathbf{M}^{(j)}$, $j = 1, 2$ and the matrix of residual variances $\mathbf{T}$

$$\mathbf{M}^{(1)} = \begin{pmatrix} 0 & \overline{\ln\frac{\mathbf{x}_1}{\mathbf{x}_2}} & \overline{\ln\frac{\mathbf{x}_1}{\mathbf{x}_3}} & \cdots & \overline{\ln\frac{\mathbf{x}_1}{\mathbf{x}_D}} \\ \overline{\ln\frac{\mathbf{x}_2}{\mathbf{x}_1}} & 0 & \overline{\ln\frac{\mathbf{x}_2}{\mathbf{x}_3}} & \cdots & \overline{\ln\frac{\mathbf{x}_2}{\mathbf{x}_D}} \\ \vdots & \vdots & \vdots & \vdots & \vdots \\ \overline{\ln\frac{\mathbf{x}_D}{\mathbf{x}_1}} & \overline{\ln\frac{\mathbf{x}_D}{\mathbf{x}_2}} & \overline{\ln\frac{\mathbf{x}_D}{\mathbf{x}_3}} & \cdots & 0 \end{pmatrix}, \mathbf{M}^{(2)} = \begin{pmatrix} 0 & \overline{\ln\frac{\mathbf{y}_1}{\mathbf{y}_2}} & \overline{\ln\frac{\mathbf{y}_1}{\mathbf{y}_3}} & \cdots & \overline{\ln\frac{\mathbf{y}_1}{\mathbf{y}_D}} \\ \overline{\ln\frac{\mathbf{y}_2}{\mathbf{y}_1}} & 0 & \overline{\ln\frac{\mathbf{y}_2}{\mathbf{y}_3}} & \cdots & \overline{\ln\frac{\mathbf{y}_2}{\mathbf{y}_D}} \\ \vdots & \vdots & \vdots & \vdots & \vdots \\ \overline{\ln\frac{\mathbf{y}_D}{\mathbf{y}_1}} & \overline{\ln\frac{\mathbf{y}_D}{\mathbf{y}_2}} & \overline{\ln\frac{\mathbf{y}_D}{\mathbf{y}_3}} & \cdots & 0 \end{pmatrix},$$

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

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

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

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

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

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

  – $\widehat{\sigma}_{rs}^2$ $r, s = 1, \dots, D$ 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).

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

  ⇒ $\mathbf{T}$ is symmetrical matrix.

ILUSTRATIVE EXAMPLE

We consider the white blood cells data set of 30 samples obtained by two different methods: microscopic inspection and image analysis. White blood cell compositional data consists of three parts: granulocytes (= part x_1), lymphocytes (= part x_2) and monocytes (= part x_3), [1].

- Data fulfill the normality assumption (Shapiro-Wilk).
- 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

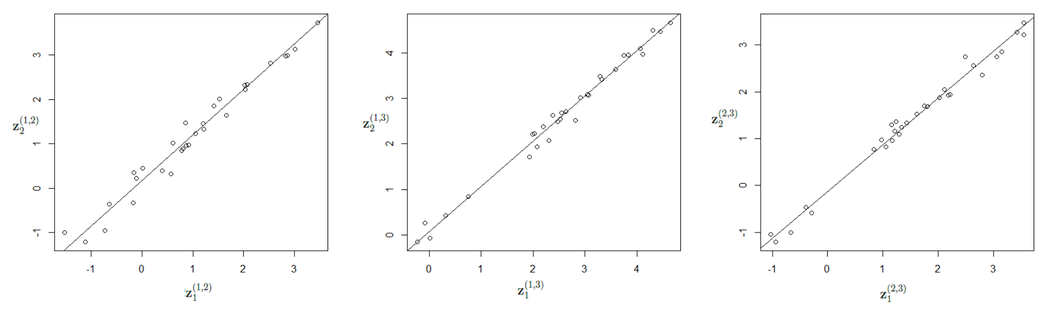

Figure 1. White blood cells data in orthonormal coordinates corresponding to $k = 1, 2, 3$ respectively, together with their propriate calibration line.

- Testing hypothesis [3]:

  – Both methods measure with the same precision of 0.2, i.e. H_0: $\sigma_{rs}^2 = 0.2^2$ v.s. H_A: $\sigma_{rs}^2 \neq 0.2^2$.

$$* \text{ Under } \mathrm{H}_0\text{: } \widehat{\sigma}_{rs}^2 \frac{n-2}{\sigma_{rs}^2} \sim \chi_{n-2}^2.$$

  ⇒ 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. H_0: $\mu_{rs} = \nu_{rs}$ v.s. H_A: $\mu_{rs} \neq \nu_{rs}$.

$$* \text{ Under } \mathrm{H}_0\text{: } T = \frac{\overline{\ln\frac{\mathbf{x}_r}{\mathbf{x}_r}} - \overline{\ln\frac{\mathbf{y}_r}{\mathbf{y}_s}} - (\mu_{rs} - \nu_{rs})}{\sqrt{(n-1)\mathrm{s}^2_{\ln\frac{\mathbf{x}_r}{\mathbf{x}_s}} + (n-1)\mathrm{s}^2_{\ln\frac{\mathbf{y}_r}{\mathbf{y}_s}}}} \sqrt{n(n-1)} \sim \mathrm{t}_{2(n-1)},$$

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

  ⇒ 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.