A compositional regression with lost values

Original problem: Blatt, Middleton and Murray (1972) published a plot where they conveyed ”the probable relationship between grain size and detrital fragment composition, based on the limited data currently available”. The plot gives, for the range of grain sizes occurring in nature, the composition on five parts (rock fragments, poly-crystalline quartz, mono-crystalline quartz, feldspar and mica). Grain size is given in ϕ scale, corresponding to the binary log of the inverse diameter. Our goal here is to fit an Aitchison (1986) trend to this data, by means of a regression. However, this composition has lots of zeroes, as can be seen in figure 1.

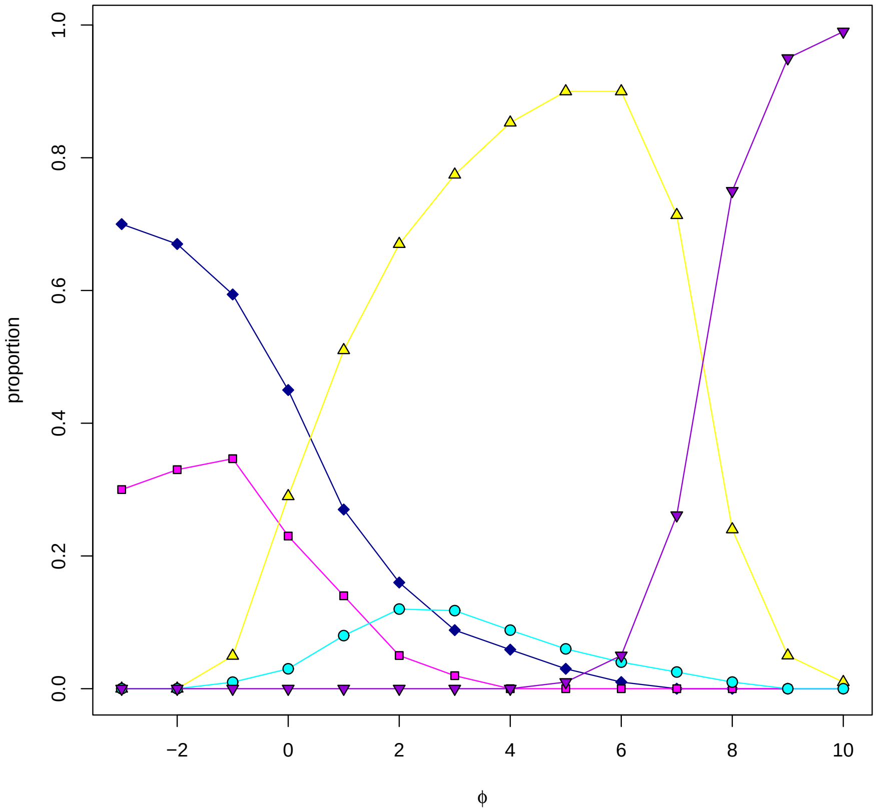

Figure 1: Data set sampled from the original figure published by Blatt et al. (1972, p. 301). Legend: rock fragments (dark blue diamonds), poly-crystalline quartz (pink squares), mono-crystalline quartz (yellow upward-triangles), feldspar (cyan circles) and mica (dark violet downward-triangles).

0.Notation: Let the grain size be represented as $\phi = -log_2 d$ (diameter of the grains). Let $\bar{\phi}$ and σ^2_ϕ be the mean and variance of ϕ (as classically defined). Let the 5-part composition $[R_f, Q_p, Q_m, F, M]$ be denoted by $\mathbf{X}$, and let $\bar{\mathbf{x}}$ be the geometric mean of $\mathbf{X}$, as corresponds to a composition in the Aitchison (1986) framework. Recall that the sample space of $\mathbf{X}$ is the $D = 5$-part simplex $\mathbb{S}^D$, which can be given an Euclidean space structure (Billheimer et al, 2001; Pawlowsy-Glahn and Egozcue, 2001) with the following operations: the Abelian group operation, called *perturbation*, is given by $\mathbf{x} \oplus \mathbf{y} = \mathcal{C}(x_1 y_1, \ldots, x_D y_D)^t$; the scalar multiplication, or *powering*, by $\alpha \odot \mathbf{x} = \mathcal{C}(x_1^\alpha, \ldots, x_D^\alpha)^t$; and the *inner product* and associated *distance* by

$$\langle \mathbf{x}, \mathbf{y} \rangle_A = \sum_{i=1}^{D} \mathrm{clr}(\mathbf{x}) \cdot \mathrm{clr}(\mathbf{y}) \,, \quad d_A^2(\mathbf{x}, \mathbf{y}) = \frac{1}{D} \sum_{i<j} \left(\ln \frac{x_i}{x_j} - \ln \frac{y_i}{y_j} \right)^2 ,$$

where $\mathrm{clr}(\mathbf{x}) = \ln \frac{\mathbf{x}}{\sqrt[D]{x_1 \cdots x_D}}$, being $\mathbf{x}, \mathbf{y} \in \mathbb{S}^D$, $\alpha \in \mathbb{R}$, the operation $\mathcal{C}(\cdot)$ closes its argument to total sum one, and the superindex t marks transposition (Aitchison, 1984; 1986).

1.Posing the problem: We look for a regression trend

$$\mathbf{x}_\phi = \mathbf{x}_0 \oplus \phi \odot \Delta\mathbf{x}, \tag{1}$$

where $\mathbf{x}_0$ is the ordinate at the origin and $\Delta\mathbf{x}$ the slope, both compositions from the same simplex as $\mathbf{x}$. Alternatively, we may look for an affine linear application

$$\mathbf{x}_\phi = \mathbf{x}_0 \oplus B\phi, \tag{2}$$

where $B : \mathbb{R} \to \mathbb{S}^D$ a linear application. Of course, these two definitions are equivalent, if we take $B\phi = \phi \odot \Delta\mathbf{x}$, in words: the image of B is a vector of $\mathbb{S}^D$ (characteristic of B) scaled with ϕ.

2.The coordinate approach: Following Pawlowsky-Glahn (2003), the Euclidean structure can be taken into account by working on the coordinates of $\mathbf{X}$ in a given basis of $\mathbb{S}^D$. Every basis can be identified with a set of exhaustive $(D-1)$ log-ratios. Since there a lots of zeroes in the data set (figure 1), some log-ratios cannot be computed. Therefore, they have to be chosen carefully. The proposed basis is

$$\mathbf{e}_1 = \mathcal{C}\left(e, 1, 1, 1, 1\right) \quad \mathbf{e}_2 = \mathcal{C}\left(1, 1, 1, e, 1\right) \quad \mathbf{e}_3 = \mathcal{C}\left(1, 1, 1, 1, e\right) \quad \mathbf{e}_4 = \mathcal{C}\left(e, e, 1, 1, 1\right)$$

because the coordinates with respect to this basis can be computed with the following set of log-ratios

$$\xi_1 = \ln \frac{R_f}{Q_p} \quad \xi_2 = \ln \frac{F}{Q_m} \quad \xi_3 = \ln \frac{M}{Q_m} \quad \xi_4 = \ln \frac{Q_p}{Q_m}.$$

Note that these log-ratios are chosen to maximize the number of cases in which both the numerator and the denominator were observed.

A classical regression model of each coordinate $\xi_i = \alpha_i + \beta_i \cdot \phi$ is straightforward to implement. The resulting parameters can be applied to the basis, e.g. $\mathbf{a} = \alpha_1 \odot \mathbf{e}_1 \oplus \alpha_2 \odot \mathbf{e}_2 \oplus \alpha_3 \odot \mathbf{e}_3 \oplus \alpha_4 \odot \mathbf{e}_4$ to recover a composition. In this way, one can estimate $\mathbf{x}_0 = \mathbf{a}$ and $\Delta\mathbf{x} = \mathbf{b}$ (Daunis-i-Estadella et al, 2002). The same application can be done with the predicted values of the regression, thus predicting the composition $\mathbf{x}_\phi$ for every ϕ (figure 2)

3.Defining a covariance: Alternatively, we can follow Eaton (1983) and estimate the parameters of Eq. (2) with a *new* covariance of $\mathbf{x}$ on ϕ, denoted $\Sigma_{x\phi} : \mathbb{R} \to \mathbb{S}^D$, and defined as the linear application fulfilling

$$\langle \mathbf{y}, \Sigma_{x\phi}\lambda \rangle_A = \mathrm{E}\left[\langle \mathbf{y}, \mathbf{x} \ominus \bar{\mathbf{x}} \rangle_A \cdot \langle \lambda, \phi - \bar{\phi} \rangle \right], \qquad (3)$$

for any vector $\mathbf{y} \in \mathbb{S}^D$ and any scalar $\lambda \in \mathbb{R}$. In the canonical basis of $\mathbb{R}$ the matrix form of $\Sigma_{x\phi}$ becomes a composition, denoted by $\mathbf{s}_{x\phi} = \Sigma_{x\phi}(1)$; in words: it is equal to the image of the canonical basis of $\mathbb{R}$ by the function $\Sigma_{x\phi}$.

4.Solving the problem: In the case of Eq. (2), the "parameters" of the regression are (Eaton, 1983)

$$\hat{B} = \Sigma_{x\phi}\Sigma_\phi^{-1} \quad \text{and} \quad \hat{\mathbf{x}}_0 = \bar{\mathbf{x}} \ominus \hat{B}\bar{\phi},$$

with $\Sigma_\phi : \mathbb{R} \to \mathbb{R}$ a linear application, the variance of ϕ in Eaton (1983) approach. In the canonical basis of $\mathbb{R}$, the matrix form of Σ_ϕ is exactly $[\sigma_\phi^2]$. As a result, the operator $\hat{B}$, which consists of the composition of two linear functions, is expressed as

$$\begin{array}{rccccc}
\Sigma_\phi^{-1}: & \mathbb{R} & \longrightarrow & \mathbb{R} & & \\
 & \phi & & \frac{1}{\sigma_\phi^2}\phi & & \\
 & & \Sigma_{x\phi}: & \mathbb{R} & \longrightarrow & \mathbb{S}^D \\
 & & & \lambda & & \lambda \odot \mathbf{s}_{x\phi} \\
B: & \mathbb{R} & \longrightarrow & \longrightarrow & \longrightarrow & \mathbb{S}^D \\
 & \phi & & & & \left(\frac{1}{\sigma_\phi^2}\phi\right) \odot \mathbf{s}_{x\phi}.
\end{array}$$

Therefore, in Eq. (1), the increment vector becomes $\Delta\mathbf{x} = \frac{1}{\sigma_\phi^2} \odot \mathbf{s}_{x\phi}$, thanks to the linear properties of $\odot$.

5.Estimating the covariance (without lost values): Developing Eq. (3) by introducing the definition of scalar products, one finds

$$\begin{aligned}
\langle \mathbf{y}, \Sigma_{x\phi}\lambda \rangle_A &= \mathrm{E}\left[\lambda(\phi - \bar{\phi})\sum_{i=1}^{D}\mathrm{clr}_i(\mathbf{y})(\mathrm{clr}_i(\mathbf{x}) - \mathrm{clr}_i(\bar{\mathbf{x}}))\right] \\
\langle \mathbf{y}, \lambda \odot \mathbf{s}_{x\phi} \rangle_A &= \sum_{i=1}^{D}\lambda\mathrm{clr}_i(\mathbf{y})\mathrm{E}\left[(\phi - \bar{\phi})(\mathrm{clr}_i(\mathbf{x}) - \mathrm{clr}_i(\bar{\mathbf{x}}))\right] \\
\lambda \cdot \langle \mathbf{y}, \mathbf{s}_{x\phi} \rangle_A &= \lambda\sum_{i=1}^{D}\mathrm{clr}_i(\mathbf{y})\mathrm{Cov}\left[\phi, \mathrm{clr}_i(\mathbf{x})\right] \\
\lambda \cdot \sum_{i=1}^{D}\mathrm{clr}_i(\mathbf{y}) \cdot \mathrm{clr}_i(\mathbf{s}_{x\phi}) &= \lambda\sum_{i=1}^{D}\mathrm{clr}_i(\mathbf{y})\mathrm{Cov}\left[\phi, \mathrm{clr}_i(\mathbf{x})\right].
\end{aligned}$$

Since this has to be satisfied for any $\lambda \in \mathbb{R}$ and any $\mathbf{y} \in \mathbb{S}^D$, then for $i = 1, \ldots, D$

$$\mathrm{clr}_i(\mathbf{s}_{x\phi}) = \mathrm{Cov}\left[\phi, \mathrm{clr}_i(\mathbf{x})\right],$$

which, being a covariance between real variables, can be estimated by standard techniques. As a curiosity, following the principle of working on coordinates of Pawlowsky-Glahn (2003) we could write $\mathbf{s}_{x\phi} = \mathrm{Cov}_A\left[\phi, \mathbf{x}, \right]$.

6.Estimating the covariance (with lost values): Recall that

$$\mathrm{clr}_i(\mathbf{s}_{x\phi}) = \mathrm{Cov}\left[\phi \cdot \mathrm{clr}_i(\mathbf{x})\right] = \mathrm{E}\left[\phi, \mathrm{clr}_i(\mathbf{x})\right] - \mathrm{E}\left[\phi\right]\mathrm{E}\left[\mathrm{clr}_i(\mathbf{x})\right].$$

This implies that, if $\mathrm{E}\left[\phi\right] = 0$, then

$$\mathrm{clr}_i(\mathbf{s}_{x\phi}) = \mathrm{E}\left[\phi \cdot \mathrm{clr}_i(\mathbf{x})\right] = \mathrm{E}\left[\mathrm{clr}_i(\mathbf{x}^{\phi})\right].$$

This simplification can be done without problem, because ϕ is in our problem fully observed, and therefore we can easily center it. Then, we may estimate

$$\mathrm{clr}_i(\mathbf{s}_{x\phi}) = \mathrm{E}\left[\mathrm{clr}_i(\phi \odot \mathbf{x})\right]$$

with the theory for estimation of mean compositions in the presence of lost values of v.d.Boogaart et al. (2006).

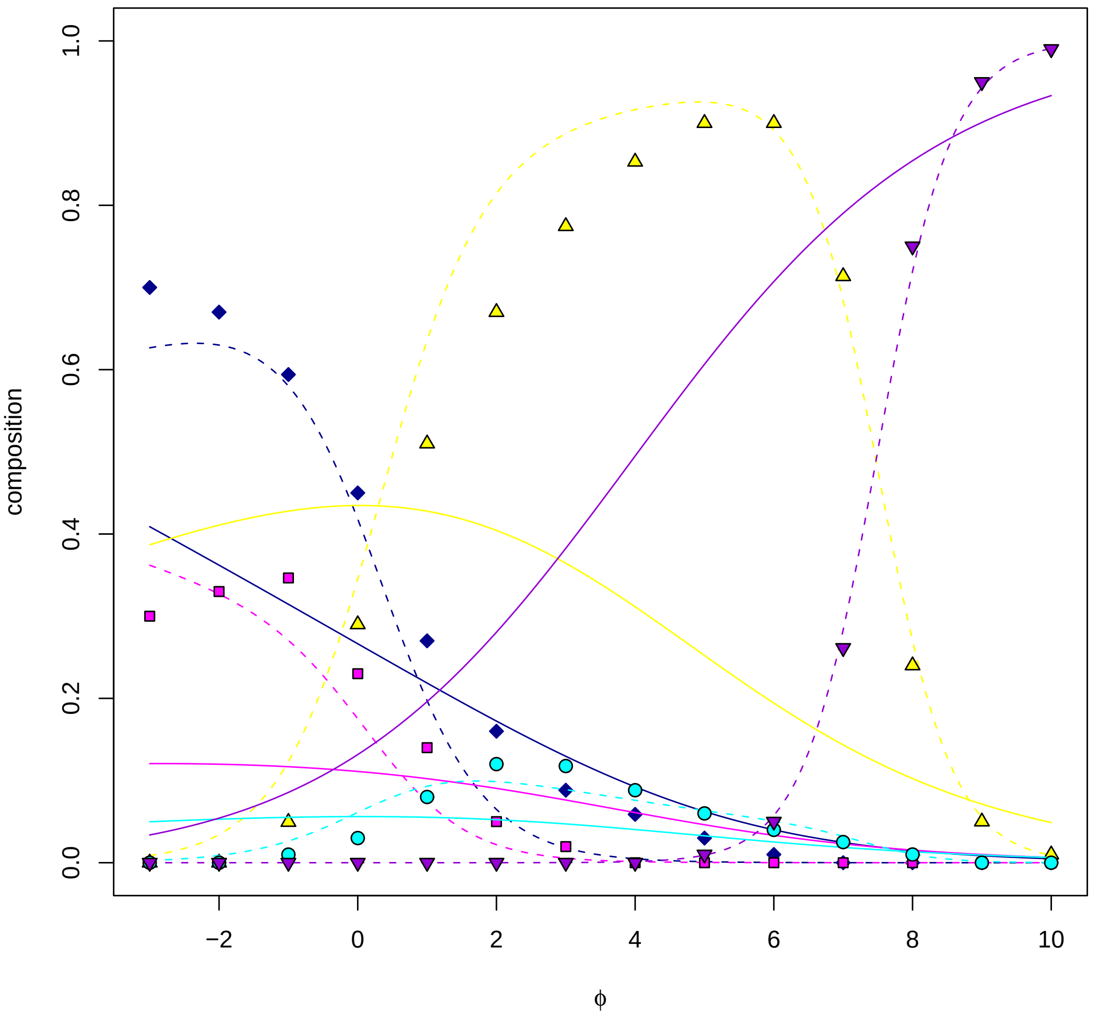

Figure 2: Result of the analysis: with the clr-projection technique (solid lines), with an ad-hoc basis (alr-type) and estimating the coefficients separately (dashed lines). Legend: rock fragments (dark blue diamonds), poly-crystalline quartz (pink squares), mono-crystalline quartz (yellow upward-triangles), feldspar (cyan circles) and mica (dark violet downward-triangles).

7.R Code:

```
# read data
 x = read.table("sedglobalbudget.txt",header=T)
# take grain size, compute its statistics, center it
  phi=x[,1]
  mphi = mean(phi)
  varphi = var(phi)
  phi = phi - mphi
# take composition, compute its mean (with losts)
  xcomp = acomp(x[,-1])
  mx = mean(xcomp)
# retrieve the projection matrix used in the mean computation
  dd = sumMissingProjector(xcomp) # as a curiosity, not needed
# compute the covariance
  auxx = phi * xcomp
  covxphi = mean(auxx)
# compute the regression coefficients (using acomp arithmetic)
 slope = (1/varphi) * covxphi
 ordinate = mx - mphi * slope
```

Results are plotted in figure 2.

NOTE: this code needs the library “compositions” in its version for handling zeroes, as can be downloaded from http://www.stat.boogaart.de/compositions/.

8.Discussion: Figure 2 shows that the fit of the curves obtained in section 2 is really good, with the single exception of rock fragments and polycrystalline quartz for low ϕ values. However, it is much better than the result one would obtain with classical statistics applied to $\mathbf{X}$ or to its standardized version.

Contrarily, the result of sections 3 to 6 is unfortunately poor. A possible explanation to this unexpected, undesirable, behavior might be related to the fact that the clr-projection approach completely ignores that lost values are ”small”: they are all considered MAR lost values, while really being BDL. Contrarily, in the simpler approach the basis has been selected so that the maximum number of zeroes ”go together”: since the ratio of two BDL’s is nearer to a true MAR, estimation with these ratios are less affected by the fact that BDL’s are not MAR. Further research is needed to adequately characterize BDLs and look for unbiased methods of estimate statistics under their presence.

References

Aitchison, J. (1984). Reducing the dimensionality of compositional data sets. *Mathematical Geology 16*(6), 617–636.

Aitchison, J. (1986). *The Statistical Analysis of Compositional Data.* Monographs on Statistics and Applied Probability. Chapman & Hall Ltd., London (UK). (Reprinted in 2003 with additional material by The Blackburn Press). 416 p.

Billheimer, D., P. Guttorp, and W. Fagan (2001). Statistical interpretation of species composition. *Journal of the American Statistical Association 96*(456), 1205–1214.

Blatt, H.; G. Middleton and R. Murray. (1972). *Origin of sedimentary rocks.* Prentice Hall Inc., New Jersey (USA). 634 p.

van den Boogaart, K.G., R. Tolosana, and M. Bren (2006). Concepts for handling of zeroes and missing values in compositional data. In Pirard, E., A Dassargues and H.B. Havenith (Eds.), *Proceedings of IAMG'06 — The 11th annual conference of the International Association for Mathematical Geology*, GeoMaC, University of Liège. CD-ROM version.

Daunis-i-Estadella, J., J. J. Egozcue, and V. Pawlowsky-Glahn (2002). Least squares regression in the simplex. In U. Bayer, H. Burger, and W. Skala (Eds.), *Proceedings of IAMG'02 — The eigth annual conference of the International Association for Mathematical Geology*, Volume I and II, pp. 411–416. Selbstverlag der Alfred-Wegener-Stiftung, Berlin, 1106 p.

Eaton, M. L. (1983). *Multivariate Statistics. A Vector Space Approach.* John Wiley & Sons.

Pawlowsky-Glahn, V. (2003). Statistical modelling on coordinates. In S. Thió-Henestrosa and J. A. Martín-Fernández (Eds.), *Compositional Data Analysis Workshop – CoDaWork'03, Proceedings.* Universitat de Girona, ISBN 84-8458-111-X, http://ima.udg.es/Activitats/CoDaWork03/.

Pawlowsky-Glahn, V. and J. J. Egozcue (2001). Geometric approach to statistical analysis on the simplex. *Stochastic Environmental Research and Risk Assessment (SERRA) 15*(5), 384–398.