Correction

Correction: Vermeesch, P. Exploratory Analysis of Provenance Data Using R and the Provenance Package. *Minerals* 2019, *9*, 193

Pieter Vermeesch [ORCID]

University College London, Gower Street, London WC1E 6BT, UK; p.vermeesch@ucl.ac.uk

There was an error in the original article [1]. Section 8 of the paper uses a simple toy example to illustrate Principal Component Analysis (PCA) of compositional data, using the additive (alr) and centred (clr) logratio transformations. It explains that the clr transformation is preferred because it produces more easily interpretable biplots. However, it failed to mention that, even though the alr- and clr-based PCA configurations of the toy example look similar, they are not the same. This is because the alr transformation is not isometric. Distances in alr-space depend on the choice of denominator, which spells trouble for PCA. As explained in Section 7 of the paper, PCA is a special case of Multidimensional Scaling (MDS). MDS is based on dissimilarity matrices, so if distances are not well defined, then neither are the MDS configuration and, hence, the principal components. The clr transformation fixes this issue.

A correction has been made to Section 8 and Figure 5 to clarify this point. It replaces the text from "Consider the following" to "in this context." with:

Consider the following trivariate (a, b and c) dataset of three (1, 2 and 3) compositions that are constrained to a constant sum ($a_i + b_i + c_i = 100\%$ for $1 \leq i \leq 3$, Figure 5):

$$X = \begin{matrix} 1 \\ 2 \\ 3 \end{matrix} \begin{matrix} a & b & c \end{matrix} \begin{bmatrix} 0.034 & 99.88 & 0.091 \\ 69.45 & 25.55 & 5.01 \\ 72.44 & 26.65 & 0.92 \end{bmatrix} \tag{12}$$

It would be wrong to apply conventional PCA to this dataset, because this would ignore the constant sum constraint. As was discussed in Section 6, PCA begins by 'centering' the data via the arithmetic mean. Section 3 showed that this yields incorrect results for compositional data. Although the *additive logratio transformation* (alr) of Equation (1) solves the closure problem, it is not suitable for PCA because it is not isometric. For example, the alr-distance between samples 2 and 3 is 1.74 if b is used as a common denominator, but 2.46 if c is used as a common denominator.

The fact that distances are not unequivocally defined in alr-space spells trouble for PCA. Recall the equivalence of PCA and classical MDS, which was discussed in Section 7. MDS is based on dissimilarity matrices, so if distances are not well defined then neither are the MDS configuration and, hence, the principal components. This issue can be solved by the *centred logratio transformation* (clr):

$$u_i = \ln\left[\frac{x_i}{g_i}\right], v_i = \ln\left[\frac{y_i}{g_i}\right], \text{ and } w_i = \ln\left[\frac{z_i}{g_i}\right] \tag{13}$$

where g_i is the geometric mean of the ith sample:

$$g_i = \exp\left[\frac{\ln[x_i] + \ln[y_i] + \ln[z_i]}{3}\right]$$

Applying the clr-transformation to the data of Equation (12) yields a new trivariate dataset:

$$X_c = \begin{array}{c} 1 \\ 2 \\ 3 \end{array} \begin{array}{ccc} \ln(a/g) & \ln(b/g) & \ln(c/g) \\ \begin{bmatrix} -3 & 5 & -2 \\ 1.21 & 0.21 & -1.42 \\ 1.79 & 0.79 & -2.58 \end{bmatrix} \end{array} \tag{14}$$

where g stands for the geometric mean of each row. Note that each of the rows of X_c adds up to zero. Thanks to the symmetry of the clr-coordinates, the distances between the rows (which are also known as Aitchison distances) are well defined. Subjecting Equation (14) to the same matrix decomposition as Equation (8) yields:

$$X_c = \begin{bmatrix} 1 \\ 1 \\ 1 \end{bmatrix} \begin{bmatrix} 0 & 2 & -2 \end{bmatrix} + \begin{bmatrix} -1.15 & 0 & 0 \\ 0.58 & -1 & 0 \\ 0.58 & 1 & 0 \end{bmatrix} \begin{bmatrix} 3.67 & 0 & 0 \\ 0 & 0.71 & 0 \\ 0 & 0 & 0 \end{bmatrix} \begin{bmatrix} 0.71 & -0.71 & 0 \\ 0.71 & 0.71 & -0.82 \\ -0.58 & -0.58 & -0.58 \end{bmatrix} \tag{15}$$

so that

$$P = \begin{bmatrix} -4.24 & 0 & 0 \\ 2.12 & -0.71 & 0 \\ 2.12 & 0.71 & 0 \end{bmatrix} \text{ and } L = \begin{bmatrix} 2.59 & -2.59 & 0 \\ 0.29 & 0.29 & -0.58 \\ 0 & 0 & 0 \end{bmatrix} \tag{16}$$

Note that, even though this yields three principal components instead two, the variance of the third component in matrix V is zero. Therefore, all the information is contained in the first two components. Furthermore, note that the first two principal components of the compositional dataset are identical to those of the PCA example shown in Section 6 (Equation (9)). This is, of course, intentional.
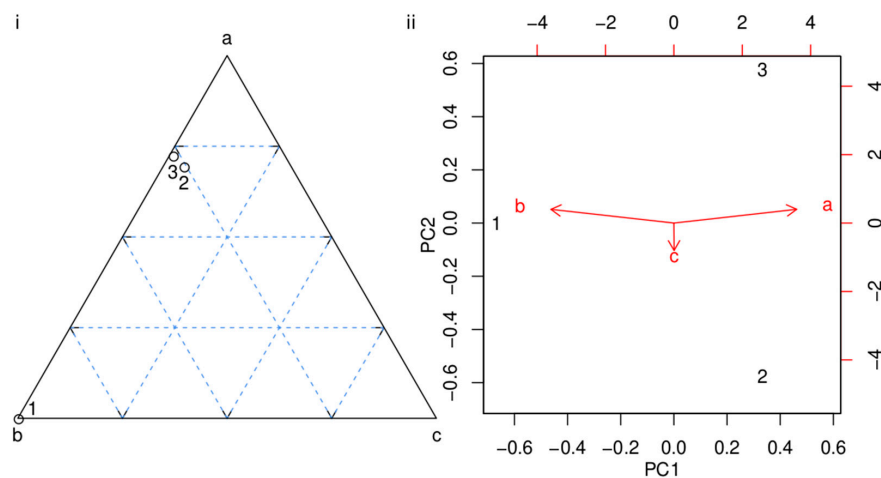
Figure 5. (**i**)—the compositional dataset of Equation (14) shown on a ternary diagram; (**ii**)—principal component biplot of the same data after centred logratio (clr) transformation.

A second correction was made to page 6, where "geometric mean" should be replaced with "logratio mean". The corrected paragraph appears below:

"5. Compute the logratio mean composition and add it to the existing ternary diagram as a red square:"

"**Figure 1.** Graphical output of Section 3. Black circles mark 20 synthetic Al_2O_3, (CaO + Na_2O) and K_2O compositions, drawn from a logistic normal distribution. The blue square marks the arithmetic mean, which falls outside the data cloud. The blue polygon marks a 2-σ confidence polygon, which plots outside the ternary diagram, in physically impossible negative space. The red square represents the logratio mean, which firmly plots inside the data cloud. The red confidence envelope marks a 95% confidence region calculated using Aitchison's logratio approach. This confidence envelope neatly fits inside the ternary diagram and tightly hugs the data."

The author states that the scientific conclusions of the remainder of the paper are unaffected. This correction was approved by the Academic Editor. The original publication has also been updated.

Reference

1.   Vermeesch, P. Exploratory Analysis of Provenance Data Using R and the Provenance Package. *Minerals* **2019**, *9*, 193. [CrossRef]