Tests for identifying the unchanging reference component of compositional data using the properties of the coefficient of variation

T. OHTA[1], H. ARAI[2] and A. NODA[3]

[1]Department of Earth Sciences, Faculty of Education and Integrated Arts and Sciences, Waseda University, Japan, tohta@waseda.jp

[2]Advanced Research Center for Science and Engineering, Waseda University, Japan

[3]Geological Survey of Japan, National institute of Advanced Industrial Sciences and Technology, Japan

Abstract

In analyses of compositional data, it is important to select a suitable unchanging component as a reference to detect the behavior of a single variable in isolation. This paper introduces two tests for detecting the unchanging component, based on a new approach that utilizes the coefficient of variation of component ratios. That is, the coefficient of variation of a compositional ratio is subject to change when the unchanging component is switched between the denominator and numerator, and the coefficient of variation tends to be small when the unchanging component occurs as the denominator against any arbitrary components (Test 1). In addition, the ratio of the component pair that gives the lowest coefficient of variation is most likely to represent the two unchanging components (Test 2). However, Tests 1 and 2 are not necessary and sufficient conditions for uniquely finding the unchanging component. To verify the effectiveness of the tests, 500 artificial data sets were analyzed and the results suggest that the tests are able to identify the unchanging component, although Test 1 underperforms when the data set includes a component with skewness greater than 0.5, and Test 2 fails when the data set includes components with a correlation coefficient greater than 0.75. These defects can be overcome by interpreting the two test results in a complementary manner. The proposed tests provide powerful yet simple criteria for identifying the unchanging component in compositional data; however, the reliability of this approach needs to be assessed in further studies.

Keywords: Compositional data, closed data, unchanging component, reference frame, coefficient of variation

Egozcue, J.J., Tolosana-Delgado, R. and Ortego, M.I. (eds.)
ISBN: 978-84-87867-76-7

1. Introduction

In the field of geosciences, scientists seek to compare compositional data to identify differences between data sets or to detect a trend within the data. An innovative approach proposed by Aitchison (1986) and many successive works made this difficult task possible by mapping the simplex sample space to the Euclidean sample space (logratio analysis) or by building a proper algebra and distance function within the simplex (stay-in-the-simplex method).

The basic problem encountered in analyzing compositional data is that they only retain relative information, which is best extracted using Aitchison's (1986) method. However, in some cases, geoscientists need to determine the behavior of a single variable in isolation; e.g., detecting gains or losses of a single element during geological processes such as magmatism, metamorphism or alteration. Four methods have been proposed for this purpose: (1) mass-balance calculations (Brimhall and Dietrich, 1987; Brimhall et al., 1988), (2) the isocon method (Grant, 1986), (3) the logratio f-value method (Woronow and Love, 1990) and (4) Pearce element ratios (Pearce, 1968).

For example, the mass-balance approach combines the chemical and physical properties of the host rock and weathered derivatives to determine the enrichment or depletion of a single element during weathering. The basic mass conservation of any element during weathering can be stated based on the volumes, densities, compositions of the host rock and weathered material as well as the mass flux between them (e.g., eq. 1 of Brimhall and Dietrich, 1987). The compositions and densities can be measured directly, and in the case of an element whose absolute abundance remained unchanged during weathering, the mass flux term can be eliminated. Subsequently, the only remaining unknown quantity, the volumetric change, can be calculated. By importing this quantity to the mass conservation equations of other elements, the absolute mass flux of a given element can be estimated (Brimhall and Dietrich, 1987; Brimhall et al., 1988). The logical operation of the isocon method is similar to this mass-balance approach (see Grant, 1986, 2005).

All four of above methods utilize the unchanging component to determine the fractionation of a single variable. However, a problem arises when the unchanging component is not self-evident, which happens to be the case in many practical data sets in geosciences. Woronow and Love (1990) and Schedl (1998) made an advance in determining a truly unchanging reference component by providing robust statistical criteria for the identification of the unchanging component in compositional data. Although these statistical methods are mathematically valid, few studies have employed these methods in practical analyses of geoscientific data. This lack of uptake may reflect the fact that the methods involve a somewhat complicated series of statistical tests and require the existence of two unchanging components within the given compositional data.

Subsequent to these studies, there has been little progress. The present study aims to introduce additional criteria for detecting the unchanging component, based on a new approach that utilizes the coefficient of variation. First, we investigate the properties of the coefficient of variation in compositional data. Then, we show that the coefficient of variation of an unchanging component meets two unique conditions that can serve as new criteria in identifying the unchanging component. If this method can be

used to readily identify the unchanging component, it would then be possible to quantify the actual variation of the single component in compositional data.

2. Definitions

The nomenclature and symbols used in this paper follow those of Aitchison (1986). Basis is a non-constrained, *D*–dimensional data vector $\mathbf{w} = (w_1, \ldots, w_D)$ and w_t is the sum of the basis. Composition, $\mathbf{x} = (x_1, \ldots, x_D)$, corresponds to the constrained data vector of **w** whose sum is 1 or 100. *D* and *n* denote the number of variables and the number of cases, respectively. Note that we only give the sub-index for the variables in all data matrix symbols and that we omit using the sub-index for the cases (e.g., w_i and x_i), for the sake of simplicity. Therefore, if the summation sign ($\sum$) is used, it indicates a summation over the cases for which the sub-index is not given.

$\overline{w_i}$, s_{w_i} and V_{w_i} denote the mean, standard deviation and coefficient of variation of w_i, respectively. The standardized *k*-th moment of w_i is expressed as $\alpha_{w_i}^k$ and the product moment coefficient of w_i and w_j is expressed as r_{w_i,w_j}.

3. Prediction of compositional data using the basis

Because **x** is in a constrained form, it is impossible to identify the unchanging component solely from **x**. Conversely, the unchanging component in **w** is obvious. Therefore, we will attempt to identify the relationship between the composition and the basis by predicting V_{x_i} for any component in compositional data in terms of the statistics of the basis. If this approach is successful, it would be possible to determine the conditions under which x_i is the unchanging component.

In terms of mathematical background, this study is largely an extension of Pearson's (1897) theory. However, we also seek to improve the formula proposed by Pearson (1897) because the associated prediction error tends to be large, as discussed below.

3.1 Predicting the mean

If the deviation of w_i is denoted as $\varepsilon_{w_i} = w_i - \overline{w_i}$, the mean of the composition x_i is:

$$\begin{aligned} \overline{x_i} &= \frac{1}{n}\sum x_i = \frac{1}{n}\sum \frac{w_i}{w_t} \\ &= \frac{1}{n}\frac{\overline{w_i}}{\overline{w_t}}\sum\left(\frac{\varepsilon_{w_i}}{\overline{w_i}}+1\right)\left(\frac{\varepsilon_{w_t}}{\overline{w_t}}+1\right)^{-1} \end{aligned} \tag{1}$$

We assume that the ratio of the deviation to the mean is small, and using the McLaurin expansion (third-order approximation), we have:

$$\overline{x_i} = \frac{1}{n}\frac{\overline{w_i}}{\overline{w_t}}\sum\left(1+\frac{\varepsilon_i}{\overline{w_i}}-\frac{\varepsilon_t}{\overline{w_t}}+\frac{{\varepsilon_t}^2}{\overline{w_t}^2}+\frac{{\varepsilon_t}^3}{\overline{w_t}^3}-\frac{\varepsilon_i\varepsilon_t}{\overline{w_i}\,\overline{w_t}}+\frac{\varepsilon_i{\varepsilon_t}^2}{\overline{w_i}\,\overline{w_t}^2}\right)$$

We neglect the cubic terms, except for the skewness term, as skewness can be large depending on the distribution. Subsequently, the mean of composition can be expressed in terms of the basis statistics as follow:

$$\overline{x_i} = \frac{\overline{w_i}}{\overline{w_t}}\left(1+{V_{w_t}}^2+\alpha_{w_t}^3{V_{w_t}}^3-r_{w_i,w_t}V_{w_i}V_{w_t}\right) \tag{2}$$

3.2. Predicting standard variation

Similarly to above, the standard deviation of the composition x_i can be denoted by the basis.

$${s_{x_i}}^2 = \frac{1}{n}\sum\left(x_i-\overline{x_i}\right)^2$$

$$= \frac{1}{n}\frac{\overline{w_i}^2}{\overline{w_t}^2}\sum\left[\left(\frac{\varepsilon_{w_i}}{\overline{w_i}}+1\right)\left(\frac{\varepsilon_{w_t}}{\overline{w_t}}+1\right)^{-1}-\left(1+{V_{w_t}}^2-r_{w_i,w_t}V_{w_i}V_{w_t}\right)\right]^2 \tag{3}$$

If the McLaurin expansion of Equation (3) is performed until the fourth-order approximation, and if we neglect the products of more than four quantities (except for the kurtosis term, as kurtosis is always greater than 1.8), we have:

$${s_{x_i}}^2 = \frac{1}{n}\frac{\overline{w_i}^2}{\overline{w_t}^2}\sum\left(\frac{{\varepsilon_{w_i}}^2}{\overline{w_i}^2}+\frac{{\varepsilon_{w_t}}^2}{\overline{w_t}^2}-2\frac{{\varepsilon_{w_t}}^3}{\overline{w_t}^3}+3\frac{{\varepsilon_{w_t}}^4}{\overline{w_t}^4}-2\frac{\varepsilon_{w_i}}{\overline{w_i}}\frac{\varepsilon_{w_t}}{\overline{w_t}}\right)$$

$$s_{x_i} = \frac{\overline{w_i}}{\overline{w_t}}\left({V_{w_i}}^2+{V_{w_t}}^2-2\alpha_{w_t}^3{V_{w_t}}^3+3\alpha_{w_t}^4{V_{w_t}}^4-2r_{w_i,w_t}V_{w_i}V_{w_t}\right)^{1/2} \tag{4}$$

3.3 Predicting the coefficient of variation

Above, we obtained two equations predicting the mean and standard deviation of the composition in terms of the basis statistics. By dividing the latter by the former, we finally obtain the formula that addresses the coefficient of variation of the composition: namely,

$$V_{x_i} = \frac{\left({V_{w_i}}^2+{V_{w_t}}^2-2\alpha_{w_t}^3{V_{w_t}}^3+3\alpha_{w_t}^4{V_{w_t}}^4-2r_{w_i,w_t}V_{w_i}V_{w_t}\right)^{1/2}}{1+{V_{w_t}}^2+\alpha_{w_t}^3{V_{w_t}}^3-r_{w_i,w_t}V_{w_i}V_{w_t}} \tag{5}$$

From this equation, we can see that even if w_i is the unchanging component, V_{xi} is not equal to zero. Moreover, the component that gives the smallest coefficient of variation is not always the unchanging

component. For this reason, the unchanging component cannot be detected directly from the compositional data.

To demonstrate the accuracy of Equation (5), we generated 100 sets of data that each contains five normally distributed variables with 100 cases in each. For these 500 variables, the means were fixed to 100 and the standard deviations were randomly selected from a value ranging from 1 to 40, thereby allowing a 40 times of differences in the value of the coefficient of variation. Then, actual V_{x_i}, and V_{x_i} calculated by Equation (5) were compared against the averaged ratio of deviation to the mean ($\varepsilon_{w_i} / \overline{w_i}$) for each variable in Figure 1.

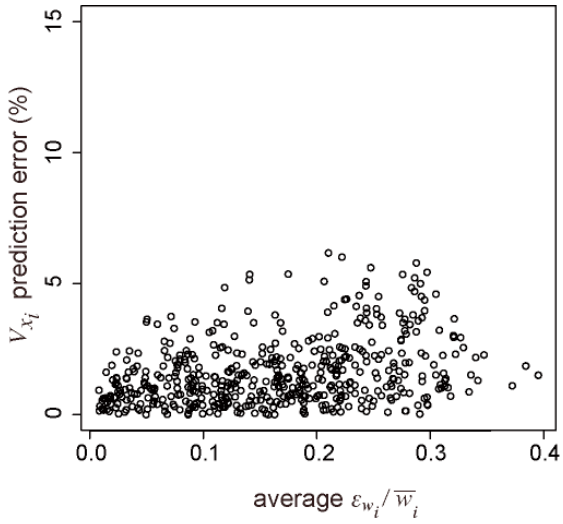

Figure 1. Prediction errors for Equations (5) in analyses of random data sets.

Although the Equation (5) was derived assuming a small value for the ratio of deviation to the mean, the prediction error does not increase significantly even for the variables with high $\varepsilon_{w_i} / \overline{w_i}$. Therefore, as long as variables do not contain significant outliers, we consider that this equation generally holds.

3.4 Coefficient of variation of ratios

If we consider the ratios of compositions, it is clear that $x_i / x_j = w_i / w_j$; thus, Equation (5) can be rewritten as follows:

$$V_{x_i/x_j} = \frac{\left({V_{w_i}}^2 + {V_{w_j}}^2 - 2\alpha_{w_j}^3 {V_{w_j}}^3 + 3\alpha_{w_j}^4 {V_{w_j}}^4 - 2r_{w_i,w_j} V_{w_i} V_{w_j} \right)^{1/2}}{1 + {V_{w_j}}^2 + \alpha_{w_j}^3 {V_{w_j}}^3 - r_{w_i,w_j} V_{w_i} V_{w_j}} \qquad (6)$$

Supposing that w_j is an unchanging component, then V_{w_j} is zero and we have $V_{x_i/x_j} = V_{w_i}$. However, if the denominator and numerator are replaced, then

$$V_{x_j/x_i} = \frac{\left(V_{w_i}{}^2 - 2\alpha_{w_i}^3 V_{w_i}{}^3 + 3\alpha_{w_i}^4 V_{w_i}{}^4\right)^{1/2}}{1 + V_{w_i}{}^2 + \alpha_{w_i}^3 V_{w_i}{}^3}$$

Accordingly, the coefficient of variation of ratios is subject to change when the unchanging component is switched between the denominator and numerator.

Figure 2 compares the values of V_{x_i/x_p} and V_{x_p/x_i} in the case that w_p is the unchanging component. The shaded area in Figure 2a shows the value range when the kurtosis of w_i is changed from 1.8 to 3.0 (from a uniform to normal distribution), and those in Figure 2b shows the value range when the skewness of w_i is changed from 0.99 to –0.99 (from positively to negatively skew-normal distributions). For Figure 2a, the value range suggests that V_{x_p/x_i} is always greater than V_{x_i/x_p}. This is also true for Figure 2b; however, the opposite relationship can occur when w_i is highly positively skewed.

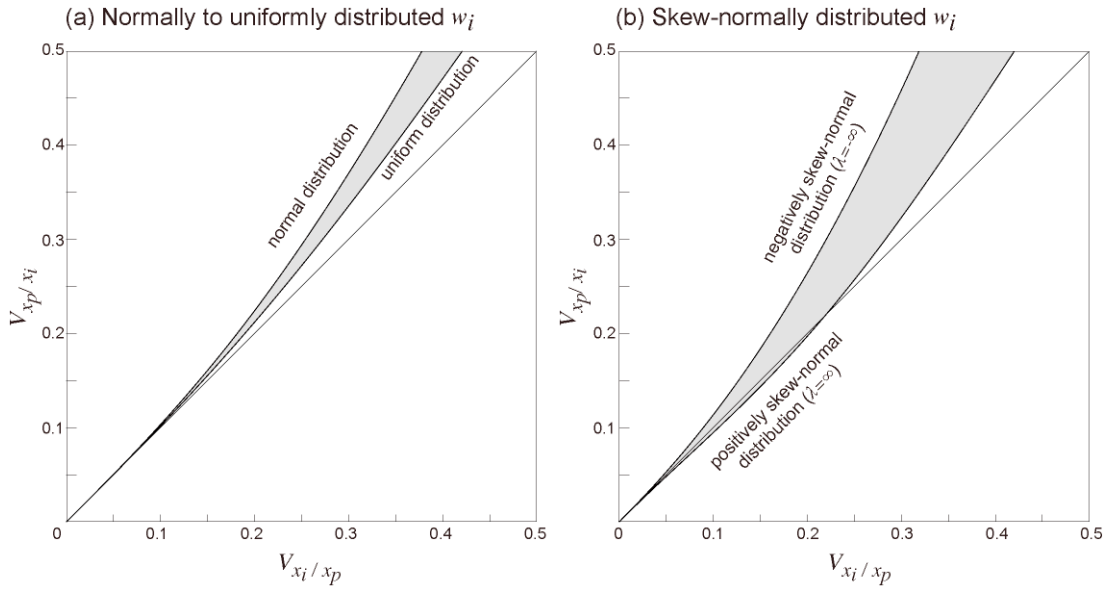

Figure 2. Comparison of the coefficients of variation of ratios when the unchanging component is the denominator (horizontal axis) and numerator (vertical axis). x_p and x_i represent the unchanging component and arbitrary component, respectively. (a) The shaded area represents the value range with changing kurtosis of the counterpart component (x_i). (b) The shaded area represents the value range with changing skewness of the counterpart component.

The above findings suggest that if w_p is the unchanging component, then for every component *i*, V_{x_i/x_p} will tend to be smaller than V_{x_p/x_i}. This relationship also holds when $V_{w_p} < V_{w_i}$, indicating that w_p need not be a completely unchanging component. Therefore, this proposition can serve as one criterion for finding the unchanging or least-changing component in compositional data (Property 1). However, Property 1 is not a necessary and sufficient condition for finding the unchanging component, as components

that give a highly positively skewed distribution can distort this condition (Fig. 2b). Consequently, additional diagnostic properties are necessary to uniquely define the unchanging component.

In the case of two unchanging components in the data ($w_p = w_q$ = constant), then $V_{w_p} = V_{w_q} = 0$ and V_{x_p/x_q} is equal to zero. Accordingly, the component pair that gives the lowest coefficient of variation of the ratio is most likely to represent the two unchanging components (Property 2). However, a problem with Property 2 is that it requires two unchanging components in the data. Another concern is that V_{x_i/x_j} can attain small values depending on the combination of V_{w_i}, V_{w_j} and r_{w_i,w_j}. In particular, in the case that $V_{w_i} = V_{w_j}$ and $r_{w_i,w_j} = 1$, that means if there are two components behaving similarly, then V_{x_i/x_j} decreases to:

$$\frac{3\alpha_{w_j}^4 {V_{w_j}}^4 - 2\alpha_{w_j}^3 {V_{w_j}}^3}{1 + \alpha_{w_j}^3 {V_{w_j}}^3}$$

Therefore, like Property 1, Property 2 is not a necessary and sufficient condition for finding the unchanging component.

4. Tests for identifying the unchanging component

As noted above, the presence of a component with a highly positively skewed distribution can distort Property 1 but not Property 2. Similarly, the presence of components that behave similarly may distort Property 2 but not Property 1. Consequently, the component that satisfies both Properties 1 and 2 is the most likely candidate for the unchanging component. On this basis, we propose two tests for identifying the unchanging component.

Test 1 is based on Property 1. First, compare the coefficients of variation for pair-wise ratios. Then, count the numbers of components for which $V_{x_i/x_p} < V_{x_p/x_i}$. If w_p is the most invariant component in **w**, then the count number would logically be D–1, where D is the number of components. Thus, the component with the largest count is the most likely candidate for the unchanging component.

Test 2 is based on Property 2. Compare the coefficients of variation for all ratio combinations. Select the four or five lowest coefficients of variation. The component that appears repeatedly in these four or five ratios is the candidate for the unchanging component. Logically, it may be sufficient to consider only the denominator of the ratio that returns the lowest coefficient of variation. However, some leeway is given in this regard because there might be a low probability of two unchanging components being simultaneously present in the same data set and a high probability of two components behaving similarly in the data set.

5. Application of the tests to artificial data sets

Five artificial data sets, each consisting of eight components (100 cases) with the last component being invariant (w_8), were generated to verify the effectiveness of Tests 1 and 2. These data sets were designed to represent extreme cases, with the intention of assessing the conditions under which the two tests are successful or unsuccessful in detecting the unchanging component (x_8 in these cases). The components of the first four data sets are designed to represent four different independent distribution classes; normal (Table 1), uniform (Table 2), negatively skew-normal (Table 3) and positively skew-normal distributions (Table 4). For these four data sets, the means of w_1 to w_7 were fixed to 100 and the standard deviations were randomly chosen from 1 to 40, thus. The fifth data set consists of normally distributed and correlated components (Table 5). The coefficient of correlation between w_1 and other variables were designed to change sequentially (0.99, 0.75, 0.30, –0.30, –0.75 and –0.99). In this case, the means and standard deviations of w_1 to w_7 were approximately 90—110 and 25—35, respectively.

Tables 1 to 5 summarize the average statistics after 100 replicated runs. The rows Test 1 and Test 2 in the tables show the number of runs for which variables x_1 to x_8 passed the tests after 100 runs. Note that these test results do not always sum up to 100 because in some of the runs, there were two or more variables simultaneously passed the test with equal counts for Tests 1 and 2.

Table 1. Statistics of normally distributed data sets (mean values after 100 runs).

BASIS	w_1	w_2	w_3	w_4	w_5	w_6	w_7	w_8
Coefficient of variation	0.353	0.294	0.253	0.207	0.156	0.111	0.059	0.000
Skewness	-0.024	-0.004	-0.005	0.004	0.000	-0.003	0.007	—
Kurtosis	2.759	2.911	2.877	2.783	2.833	2.907	2.911	—
Correlation with w_1	1.000	0.016	-0.005	0.001	0.001	0.001	-0.007	—
COMPOSITION	x_1	x_2	x_3	x_4	x_5	x_6	x_7	x_8
Coefficient of variation	0.322	0.270	0.235	0.198	0.158	0.127	0.097	0.077
Skewness	-0.194	-0.097	-0.086	0.023	0.127	0.200	0.366	0.466
Kurtosis	2.864	2.982	2.947	2.897	3.005	3.006	3.146	3.298
Correlation with x_1	1.000	-0.227	-0.257	-0.261	-0.292	-0.331	-0.374	-0.431
TEST RESULTS	x_1	x_2	x_3	x_4	x_5	x_6	x_7	x_8
Test 1	0	0	1	0	2	14	25	70
Test 2 (lowest 5 pairs)	0	0	0	0	0	0	7	93

Table 2. Statistics of uniform distributed data sets (mean values after 100 runs).

BASIS	w_1	w_2	w_3	w_4	w_5	w_6	w_7	w_8
Coefficient of variation	0.336	0.285	0.243	0.195	0.141	0.095	0.049	0.000
Skewness	0.022	0.012	0.007	-0.010	-0.040	-0.003	0.006	—
Kurtosis	1.782	1.778	1.771	1.772	1.804	1.812	1.786	—
Correlation with w_1	1.000	-0.003	0.001	-0.005	-0.017	-0.001	-0.003	—
COMPOSITION	x_1	x_2	x_3	x_4	x_5	x_6	x_7	x_8
Coefficient of variation	0.305	0.260	0.223	0.185	0.144	0.112	0.086	0.073
Skewness	-0.009	0.005	0.018	0.089	0.132	0.182	0.249	0.340
Kurtosis	1.888	1.929	1.976	2.118	2.339	2.549	2.689	2.817
Correlation with x_1	1.000	-0.246	-0.245	-0.265	-0.297	-0.334	-0.379	-0.412

TEST RESULTS	x_1	x_2	x_3	x_4	x_5	x_6	x_7	x_8
Test 1	0	1	0	2	5	22	28	62
Test 2 (lowest 5 pairs)	0	0	0	0	0	0	5	95

The data sets shown in Tables 1 and 2 represent a case in which every component is normally and uniformly distributed, respectively. For both the normal and uniform cases, the detection ability of Test 1 in selecting the true unchanging component (x_8) was better than 60%. This result may appear to be unsatisfactory, but Test 1 chose x_6 and x_7 in the cases when these were nearly invariant variables. Therefore, Test 1 detected the true or nearly unchanging components in majority of the 100 replicated runs. The detection ability of Test 2 was almost perfect (>90%; Tables 1 and 2). These results suggest that although both tests successfully detected the unchanging and/or least changing components, Test 2 generally performed better than Test 1.

Table 3. Statistics of negatively normal-skew distributed data sets (mean values after 100 runs).

BASIS	w_1	w_2	w_3	w_4	w_5	w_6	w_7	w_8
Coefficient of variation	0.190	0.173	0.174	0.162	0.169	0.154	0.154	0.000
Skewness	-0.192	-0.512	-0.611	-0.767	-0.787	-0.861	-0.879	—
Kurtosis	2.944	3.231	3.057	3.367	3.271	3.447	3.492	—
Correlation with w_1	1.000	0.004	-0.013	-0.004	0.004	-0.000	0.006	—
COMPOSITION	x_1	x_2	x_3	x_4	x_5	x_6	x_7	x_8
Coefficient of variation	0.181	0.171	0.172	0.163	0.171	0.158	0.156	0.060
Skewness	-0.045	-0.246	-0.308	-0.416	-0.411	-0.378	-0.443	0.657
Kurtosis	3.126	3.368	3.206	3.492	3.448	3.567	3.416	3.667
Correlation with x_1	1.000	-0.172	-0.194	-0.162	-0.113	-0.171	-0.161	0.002
TEST RESULTS	x_1	x_2	x_3	x_4	x_5	x_6	x_7	x_8
Test 1	2	2	0	0	0	0	0	96
Test 2 (lowest 5 pairs)	0	3	0	0	2	5	1	89

Table 3 shows the case in which the basis artificial data set has varying degrees of negative skewness (from 0.0 to –0.9) and varying coefficient of variation (from 0.01 to 0.40). Both Test 1 and 2 were almost perfect in detecting x_8 (Table 3). The performance of Test 1 was improved compared with the cases in Tables 1 and 2, because the differences between $V_{xi/x8}$ and $V_{x8/xi}$ becomes more apparent when the data are negatively skewed (see Fig. 2).

Table 4. Statistics of positively normal-skew distributed data sets (mean values after 100 runs).

BASIS	w_1	w_2	w_3	w_4	w_5	w_6	w_7	w_8
Coefficient of variation	0.161	0.119	0.117	0.117	0.105	0.109	0.102	0.000
Skewness	0.246	0.483	0.690	0.829	0.824	0.902	0.892	—
Kurtosis	3.031	3.122	3.490	3.566	3.404	3.577	3.652	—
Correlation with w_1	1.000	-0.005	0.002	-0.008	0.006	-0.014	-0.007	—

COMPOSITION	x_1	x_2	x_3	x_4	x_5	x_6	x_7	x_8
Coefficient of variation	0.151	0.115	0.113	0.111	0.103	0.106	0.100	0.046
Skewness	0.114	0.296	0.453	0.577	0.541	0.581	0.574	-0.017
Kurtosis	2.941	2.931	3.114	3.245	3.147	3.187	3.257	2.837
Correlation with x_1	1.000	-0.208	-0.204	-0.214	-0.157	-0.188	-0.184	-0.048
TEST RESULTS	x_1	x_2	x_3	x_4	x_5	x_6	x_7	x_8
Test 1	4	8	18	20	22	31	28	0
Test 2 (lowest 5 pairs)	1	0	0	1	1	0	1	96

Table 4 shows the case in which the basis has a positive skew-normal distribution with varying degrees of skewness (from 0.0 to 0.9). This is a case in which Test 1 was assumed to be ineffective in detecting the unchanging component, as expected, Test 1 failed to detect x_8; instead, strongly positively skewed variables such as x_5, x_6 and x_7 erroneously passed the test. However, in most of the 100 runs, Test 1 detected x_8 as a second candidate. That is, $V_{x_i/x_8} < V_{x_8/x_i}$ count numbers were large, but not the largest.

Table 5. Statistics of correlated normal distributed data sets (mean values after 100 runs).

BASIS	w_1	w_2	w_3	w_4	w_5	w_6	w_7	w_8
Coefficient of variation	0.298	0.289	0.294	0.298	0.299	0.297	0.298	0.000
Skewness	0.006	0.013	-0.004	-0.010	0.040	0.003	-0.001	-
Kurtosis	2.956	2.950	2.854	2.840	2.923	2.862	2.896	-
Correlation with w_1	1.000	0.990	0.747	0.285	-0.281	-0.751	-0.990	-
COMPOSITION	x_1	x_2	x_3	x_4	x_5	x_6	x_7	x_8
Coefficient of variation	0.272	0.263	0.262	0.265	0.285	0.317	0.339	0.073
Skewness	-0.126	-0.116	-0.170	-0.172	-0.006	0.166	0.278	0.441
Kurtosis	2.995	2.985	2.917	2.948	2.950	2.967	3.079	3.287
Correlation with x_1	1.000	0.987	0.680	0.098	-0.498	-0.814	-0.911	-0.309
TEST RESULTS	x_1	x_2	x_3	x_4	x_5	x_6	x_7	x_8
Test 1	0	0	0	0	0	0	0	100
Test 2 (lowest 5 pairs)	67	70	1	0	0	10	10	7

Table 5 shows normally distributed data with varying degrees of correlation with x_1 (ranging from 0.99 to –0.99). In this case, the coefficient of variation was fixed to around 0.3. Test 1 succeeded in detecting x_8 uniquely. Meanwhile, Test 2 failed to detect x_8; instead, variables with high positive correlations (x_1 and x_2) erroneously passed Test 2, as expected from the mathematical arguments.

In summary, Tests 1 and 2 appear to represent a powerful tool in detecting the unchanging component. However, the results suggest that the accuracy of Test 1 tends to decrease when there is a positively skewed component (greater than 0.5) and the result of Test 2 is doubtful when there are components with a high coefficient of correlation (greater than 0.75). We found that Test 2 returns a more robust result than Test 1 in the case when nearly unchanging component is present in the data set.

Based on these results and given that priority should be given to the results of Test 2, the

following steps are proposed for identifying the unchanging component. Step 1: In the case that both tests identify the same component, this component is chosen as the reference component. Step 2: In the case that the two tests return different results but with some common points, the first candidate is the component that passed Test 2 and that also returns a relatively large number in the result obtained using Test 1. Step 3: In the case that the results of Tests 1 and 2 are completely different, we should take into account geological and empirical knowledge in making a final decision. That is, Step 3.1: when the researcher is certain that the data set does not contain multiple unchanging components and/or does not include highly correlated components, the component identified by Test 1 is the first candidate regardless of the result of Test 2. Step 3.2: If, however, one is certain that the data set includes multiple unchanging components and/or if the data set definitely includes a highly positively skewed component, priority should be given to the results of Test 2.

The artificial examples presented in this study indicate that the two tests represent powerful yet simple criteria for identifying the reference component in compositional data. However, given that the effectiveness of the tests was assessed based on somewhat simple cases, it is necessary to further assess the reliability of this approach based on analyses of natural and simulated data sets.

6. Conclusions

The selection of a suitable reference component is important in undertaking an analysis of a single variable in isolation for compositional data. The tests introduced in this paper represent criteria for detecting the unchanging component, based on a new approach that utilizes the coefficient of variation. The tests constitute a promising method for determining the unchanging component, as demonstrated by analyses of artificial data sets.

The performances of the tests are limited in that they are not necessary and sufficient conditions for detecting the unchanging component. This problem can be overcome by using the results of the two tests in a complementary manner. Another important aspect of the proposed approach is that the researcher should take into account geological and empirical knowledge when considering the final decision regarding identifying the unchanging component.

Another limitation is that the mathematical development of the present argument assumed a small value for the ratio of deviation to the mean, in order to conduct the McLaurin expansion. Therefore, it is highly probable that the equations introduced in this paper do not hold in the case that outliers occur in the given data. This limitation should also be investigated in further works.

Acknowledgements

We are indebted to Dr. Hilmar von Eynatten for reading and improving the manuscript. This work was financially supported by a Waseda University Grant for Special Research Project (2009B-070) and a JSPS Grant-in-Aid for Young Scientists (B) (No. 22740332; T. Ohta).

References

Aitchison J (1986) The statistical analysis of compositional data. Chapman & Hall, London, 416 p

Azzalini A (1985) A class of distributions which includes the normal ones. Scand. J. Stat. 12: 171-178

Brimhall GH, Dietrich WE (1987) Constitutive mass balance relations between chemical-composition, volume, density, porosity, and strain in metasomatic hydrochemical systems - Results on weathering and pedogenesis. Geochim Cosmochim Acta 51: 567-587

Brimhall GH, Lewis CJ, Ague JJ, Dietrich WE, Hampel J, Teague T, Rix P (1988) Metal enrichment in bauxites by deposition of chemically mature aeolian dust. Nature 333: 819-824

Gómez HW, Venegas O, Bolfarine H (2007) Skew-symmetric distributions generated by the distribution function of the normal distribution. Environmetrics 18: 395-407

Grant JA (1986) The isocon diagram - A simple solution to gresens' equation for metasomatic alteration. Econ Geol 81: 1976-1982

Grant JA (2005) Isocon analysis: A brief review of the method and applications. Phys Chem Earth 30: 997-1004

Pearce TH (1968) A contribution to the theory of variation diagrams. Contrib Mineral Petrol 19: 142-157

Pearson K (1897) Mathematical contribution to the theory of evolution: on a form of spurious correlation which may arise when indices are used in the measurement of organs. Proc Royal Soc 60: 489-498

Schedl A (1998) Log ratio methods for establishing a reference frame for chemical change. J Geol 106: 211-228

Turner BF, Stallard RF, Brantley SL (2003) Investigation of in situ weathering of quartz diorite bedrock in the Rio Lcacos basin, Luquillo Experimental Forest, Puerto Rico. Chem Geol 202: 313-341

White AF, Blum AE, Schulz MS, Vivit DV, Stonestrom DA, Larsen M, Murphy SF, Eberl D (1998) Chemical weathering in a tropical watershed, Luquillo mountains, Puerto Rico: I. Long-term versus short-term weathering fluxes. Geochim Cosmochim Acta 62: 209-226

Woronow A, Love KM (1990) Quantifying and testing differences among means of compositional data suites. Math Geol 22: 837-852