Decomposing compositional data: minimum chi-squared reduced-rank approximations on the simplex

Gert Jan Weltje
Department of Applied Earth Sciences
Delft University of Technology
PO Box 5028
NL-2600 GA Delft
The Netherlands
E-mail: g.j.weltje@citg.tudelft.nl

Introduction

The logratio transformation (Aitchison, 1981, 1986) opened the way to statistically rigorous analysis of compositional data. Most statistical problems in compositional data analysis can be formulated in terms of logratios, and solved accordingly. However, some problems are more easily described and solved in terms of raw compositional variables. Raw compositions with \( k \) components are restricted to a \( k-1 \) dimensional subspace of the real space, a so-called simplex, defined by the following two constraints:

\[
\sum_{i=1}^{k} p_i = 1 \quad \text{and} \quad p_i \geq 0.
\]

Compositional variation within suites of rocks and rock-forming minerals is often believed to reflect linear mixing. The corresponding mathematical model, expressed in terms of raw compositional variables may be represented in matrix notation as:

\[
P = MB + E
\]

where \( P \) is a \( (n \times k) \) matrix of observed compositions, believed to represent a set of \( n \) mixtures, \( M \) is a \( (n \times q) \) matrix of mixing coefficients, and \( B \) a \( (q \times k) \) matrix of end members. The number of end members in the mixture is defined as \( q \). \( E \) is a \( (n \times k) \) matrix of error terms. Note that \( M \) and \( B \) are also compositions, so that the rows of \( E \) are forced to sum to zero.

In many cases of linear mixing, \( B \) is assumed to be known exactly. For instance, in a procedure known as normative partitioning, chemical compositions of rocks are expressed as mixtures of perfect model minerals, whose chemical compositions are predicted by theory. However, typical mixing problems in sedimentary geology are of an explicit nature, characterised by a near total lack of information about the end members. The nonlinear inverse mixing problem in sedimentary geology relies on the existence of an unobservable matrix of perfect mixtures, which is defined as an exact convex linear combination of matrices \( M \) and \( B \):

\[
X_q = MB
\]
where $X_q$ is a $(n \times k)$ matrix of perfect mixtures of exact rank $q$, implying that the end members are linearly independent. It thus follows that $q \leq \min(n,k)$ and:

$$P = X_q + E$$

In other words, $X_q$ is a reduced-rank approximation of the compositional data $P$. The main purpose of this contribution is to present a method for estimating $X_q$ which permits statistically rigorous testing of the goodness-of-fit.

**Current estimation methods**

The estimation methods developed in the earth sciences (Renner, 1993; Weltje, 1997) combine factorisation of $P$ by means of a singular value decomposition with constrained least squares approximation to ensure that all elements of $X_q$ are nonnegative. This numerical approach is computationally efficient, because there is as yet no need to estimate $M$ and $B$ (Weltje, 1997). Instead, use can be made of the unidentifiability (nonuniqueness) of model parameters to postulate the existence of a $(q \times q)$ transformation matrix $T$:

$$X_q = MB = AT^{-1}TF = AF$$

The estimation problem is now reduced to finding $X_q = AF$, where $F$ is a $(q \times k)$ matrix of the first $q$ transposed eigenvectors of the data matrix $P$ scaled to unit sum, and $A$ is a $(n \times q)$ matrix of coordinates onto the eigenvectors. This problem must be solved for each row of $P$ separately:

Minimise $\|F^T a^T - p^T\|_2$

Subject to $a1 = 1$ Exact linear equality constraints (unit sum)

Subject to $F^T a^T \geq h$ Exact linear inequality constraints (nonnegativity/positivity)

Where $h$ is a $k$-vector of lower bounds on $X$ values: $h \geq 0$. It follows that the elements of $X_q$ are restricted to the interval $[h_j; 1 - \sum_{j=1}^k h_j]$.

This method of reduced-rank approximation may be used to estimate the number of linearly independent end members $q$ needed to model $P$ independently of the actual end-member compositions $B$. The weakness of this approach is the lack of a formal goodness-of-fit test for $X_q$. Although tests with synthetic data provided support for the use of certain empirical measures (Weltje, 1995, 1997; Prins & Weltje, 1999a, 1999b; Van der Ark, 1999) this problem was not fully solved.

If the number of linearly independent end members has been estimated, a physically realistic mixing model of the data can be produced by solving the following set of equations:

Minimise $\|MT - A\|_2$

Subject to $1T = 1$ Exact linear equality constraints (unit sum)

Subject to $T^{-1}B = F$ Exact linear equality constraints (unit sum: nonlinear in $T$)

Subject to $B \geq 0$ Exact linear inequality constraints (nonnegativity)
Subject to  $M \geq 0$  

Exact linear inequality constraints (nonnegativity)

A detailed account of the procedure for estimating $M$ and $B$ is beyond the scope of this contribution (for details see Weltje, 1997). However, it is important to note that it may not be possible to approximate $P$ as a product of two entirely nonnegative matrices $M$ and $B$. Hence, a set of $q$ linearly independent end members may not exist, even if the fit of $X_q$ to $P$ is good. The implicit estimation of $X_q$ thus allows the detection of possible row degeneracy among the true end members $B$.

An estimation method developed for latent budget analysis (De Leeuw & Van der Heijden, 1988), an end-member model used in the social sciences, is based on the EM algorithm for augmentation of incomplete data (Dempster et al., 1977). This method employs a maximum likelihood criterion to simultaneously estimate $X_q$, $M$ and $B$, based on the assumption that $P$ follows a product-multinomial distribution. According to this model, each composition (row of $P$) is considered a random sample from an unknown multinomial distribution. Although this approach potentially allows a statistically rigorous assessment of the discrepancies between $X_q$ and $P$, it is notoriously slow to converge, and it does not allow testing of the rank of $X_q$ without estimating $M$ and $B$.

In the following sections a new estimation method will be presented that combines the advantages of the above approaches:

- Detection of row degeneracy among end members by implicit estimation of $X_q$;
- Minimisation of a statistically rigorous measure of discrepancy between $X_q$ and $P$.

**Distributional assumptions**

A statistically rigorous measure of discrepancy between $P$ and $X_q$ may be formulated by regarding compositions as enumerative data, i.e., counts of discrete units. This is strictly true for many types of geological data, such as point counts of mineral grains or map regions. However, other types of compositional data, with units of measurement not corresponding to discrete entities, may also be treated as enumerative. The "count length" (sample size) $N$ of compositional data can be estimated from the spread of a series of replicate analyses, because samples from a population with fixed composition follow a multinomial distribution if all variation is attributable to random error.

Goodness-of-fit statistics for the multinomial distribution have been studied extensively by Cressie & Read (1984) and Read & Cressie (1988), who showed that all commonly used statistics can be regarded as members of a one-parameter family of power-divergence statistics. For large values of $N$, as commonly encountered in geological data, the behaviour of all power-divergence statistics tends towards that of the well-known Pearson's chi-squared statistic (PCS). Hence, for all practical purposes, minimisation of PCS may be regarded as equivalent to maximum likelihood estimation.

Estimation of sample size $N$ from a series of $m$ replicate compositions with $k$ constituents proceeds as follows:

$$PCS = N \sum_{i=1}^{m} \sum_{j=1}^{k} \frac{(p_{ij} - \hat{p}_j)^2}{\hat{p}_j}, \text{ where } \hat{p}_j = \frac{1}{m} \sum_{i=1}^{m} p_{ij}, \text{ the vector of arithmetic means.}$$

The fit of the replicates to the model is good by definition:
\[ PCS = E \left\{ X^2 \right\} = \chi^2_{(Pr=0.5)}, \text{ with } (m-1)(k-1) \text{ degrees of freedom.} \]

By combining the above expressions we obtain:

\[ N = \frac{\chi^2_{(Pr=0.5)}}{\sum_{i=1}^{m} \sum_{j=1}^{k} (p_{ij} - \hat{p}_j)^2} \]

Strictly speaking, every matrix of compositional data follows a product-multinomial rather than a multinomial distribution, because of the constant row-sum constraint. However, the goodness-of-fit tests for both types of distributions are algebraically equivalent (Read & Cressie, 1988), provided that the number of degrees of freedom is reduced to account for this constraint. Every compositional data matrix thus may be treated as a series of samples from one or more fixed populations. In theory, each composition represents a sample that could have been drawn from a different population, but this need not be the case. The main purpose of end-member modelling is to quantify the relations between these populations. For example, if sediments have been sampled across an entire sedimentary basin fed by multiple sediment sources, each sample can be thought of as representing the local sediment population (Weltje, 2001). Each of these local populations in turn represents a mixture of a limited number of end-member populations, whose mixing proportions vary continuously in space. There are no rigid assumptions with respect to the distribution of mixing proportions on the simplex spanned by the end members.

The Minimum Chi-Squared Decomposition algorithm

The basic equation for columnwise scaling of the matrix of compositional data is:

\[ W = PG^{-1} \]

where \( G \) is a diagonal \((k \times k)\) matrix of nonnegative weights. The objective function we seek to minimise is Pearson's chi-squared statistic (PCS):

\[ O_{PCS} = N \sum_{j=1}^{k} \sum_{i=1}^{n} \frac{(w_{ij} - \hat{w}_j)^2}{\hat{w}_j} \]

\[ = N \sum_{j=1}^{k} \sum_{i=1}^{n} \frac{(p_{ij} - x_j)^2}{x_j} \]

We prefer least-squares estimation to maximum likelihood methods, because the latter are not as computationally efficient (Van der Ark, 1999). The method of constrained weighted least squares (CWLS) minimises the objective function:

\[ O_{WLS} = \sum_{j=1}^{k} \sum_{i=1}^{n} (w_{ij} - \hat{w}_j)^2 \]

\[ = \sum_{j=1}^{k} g_j^{-2} \sum_{i=1}^{n} (p_{ij} - x_j)^2 \]

subject to nonnegativity and unit-sum constraints on the estimated composition vectors \( x \).

By equating these two objective functions we obtain the following expression for the minimum chi-squared (MCS) scaling factor of the \( j \)-th column:

\[ g_j = \frac{\sum_{i=1}^{n} (p_{ij} - x_j)^2}{N \sum_{i=1}^{n} x_j^2} \]
The above expression shows that the reduced-rank estimate $X_q$ is required to calculate the MCS scaling factors, implying that they must be estimated by an iterative process. An initial guess of the MCS scaling factors is based on the fact that the case $q = 1$ has a simple solution known from chi-squared testing theory (as outlined above for the analysis of replicates). The null hypothesis implies compositional homogeneity of $P$:

\[ H_0 : q = 1 \]

\[ H_1 : q > 1 \]

The trivial mixing model with one end member that best fits the data is simply the vector of arithmetic column means. Substituting the one end-member solution into the general expression gives:

\[ g_j = \frac{\sum_{i=1}^{n} P_{ij}}{N_n} \]

The objective functions are now identical and approximately distributed as $\chi^2$ with $(n-1)(k-1)$ degrees of freedom. If the null hypothesis is not accepted, more advanced mixing models are needed to account for the variation in $P$, implying that $q > 1$.

For the purpose of estimating $X$ in case $q > 1$, we define the vector of MCS scaling factors as $\bar{g}$, the vector of diagonal elements of $G$ normalised to unit sum. In other words, we treat $\bar{g}$ as a composition because the goodness-of-fit of the model is solely determined by the relative values of its elements:

\[ \bar{g}_j = \frac{g_j}{\sum_{j=1}^{q} g_j} \]

The iterative estimation starts by scaling the data using the current estimate of MCS scaling factors:

\[ w_{ij} = \frac{P_{ij}}{\bar{g}_j} \]

Then the eigenvectors $V$ (right singular vectors) of $W$ are extracted by a singular value decomposition:

\[ W = U S V^T \]

A set of reference vectors $F$ is calculated from the first $q$ columns of $V$ in three steps:

First the column sums of vector coefficients are constrained to be nonnegative:

\[ \sum_{j=1}^{k} v_{jm} < 0 \quad \Rightarrow \quad v_{jm} = -v_{jm} \quad \forall \quad j = 1,2,\ldots, k, \quad m = 1,2,\ldots, q \]

Then a fully nonnegative matrix $\Phi$ is formed by the linear transformation $\Phi = V T$, where:

\[ T = \begin{bmatrix}
1 & 1 & 1 & \Lambda & 1 \\
0 & \alpha_{22} & 0 & \Lambda & 0 \\
0 & 0 & \alpha_{33} & 0 & M \\
M & M & O & O & 0 \\
0 & 0 & \Lambda & 0 & \alpha_{qq}
\end{bmatrix} \]

The $\alpha_{nm}$ are calculated from the coefficients of each column of $V$: 

\[ \alpha_{mn} = \max \left( I \left( \frac{v_{jm}, v_{ij}}{v_{jm}} \right) \right) \]

Where \( I \) represents the indicator function:
\[
I \{ v \geq 0 \} = 0 \\
I \{ v < 0 \} = 1
\]

The final step combines scaling and transposition to form \( F \):
\[
f_{mj} = \frac{\phi_{jm}}{\sum_{j=1}^{k} g_{j} \phi_{jm}}
\]

The reference vectors \( F \) now obey the following constraints:
\[
F \tilde{G} = 1 \quad \text{Linear equality (unit sum)} \\
F \geq 0 \quad \text{Linear inequality (nonnegativity)}
\]

An important result of the above calculations is that \( X_q = AF\tilde{G} \) with unknown coefficients \( A \) all of the same order of magnitude. It follows that \( X_q \tilde{G}^{-1} = AF \), implying that the problem of reduced-rank approximation may be formulated in unweighted form as:

Minimise \[ \| F^T a^T - w^T \|_2 \]
Subject to \( a \| 1 \| = 1 \quad \text{Linear equality (unit sum)} \\
Subject to \( F^T a^T \geq \tilde{G}^{-1} h \quad \text{Linear inequality (nonnegativity/positivity)}
\]

This is a so-called LSIE problem (least-squares with inequality and equality constraints), which may be solved by applying the following transformations:
- Transformation of a LSIE to a LSI problem by removal of linear equality constraints (LeMaitre, 1979);
- Transformation of a LSI to a LDP (least-distance programming) problem of general form:
  Minimise \[ \| y - d \|_2 \] subject to \( Ax \geq b \) (Lawson & Hanson, 1974; Menke, 1989)
- Transformation of a LDP to a NNLS (nonnegative least-squares) problem of general form:
  Minimise \[ \| y - d \|_2 \] subject to \( y \geq 0 \) (Lawson & Hanson, 1974; Menke, 1989)

The solution to the LSIE problem is obtained by application of the inverse transformations to the NNLS solution. This completes the reduced-rank estimation for the current \( \tilde{g} \).

The vector of MCS scaling coefficients is estimated by Powell's Direction Set Method, a multidimensional minimisation method that does not require calculation of derivatives (Press et al., 1994). Multiple random starts may be used to investigate the behaviour of PCS as a function of the composition of \( \tilde{g} \). Limited experimentation has shown the presence of several local minima, which could be a matter of concern in applications to large data sets. In the data analysed thus far, multiple minima have comparable PCS values, which merely suggests that "valley floors" are not quite flat.

The resulting minimum chi-squared estimate is approximately distributed as \( \chi^2 \) with \( (n-q)(k-q) \) degrees of freedom, which allows for rigorous goodness-of-fit testing of the reduced rank approximation. The probability of the model \( X \) under the null hypothesis as estimated from
PCS is known to deviate from the theoretical $\chi^2$ in cases with low expected frequencies (Read & Cressie, 1988; Greenwood & Nikulin, 1996), which is not uncommon for geological data. In such cases an exact chi-squared test should be used (Romesburg et al., 1981).

References
