

Chapter 42

Linear Unmixing in the Geologic Sciences: More Than A Half-Century of Progress



William E. Full

Abstract For more than a half-century, scientists have been developing a tool for linear unmixing utilizing collections of algorithms and computer programs that is appropriate for many types of data commonly encountered in the geologic and other science disciplines. Applications include the analysis of particle size data, Fourier shape coefficients and related spectrum, biologic morphology and fossil assemblage information, environmental data, petrographic image analysis, unmixing igneous and metamorphic petrographic variable and the unmixing and determination of oil sources, to name a few. Each of these studies used algorithms that were designed to use data whose row sums are constant. Non-constant sum data comprise what is a larger set of data that permeates many of our sciences. Many times, these data can be modeled as mixtures even though the row sums do not sum to the same value for all samples in the data. This occurs when different quantities of one or more end-member are present in the data. Use of the constant sum approach for these data can produce confusing and inaccurate results especially when the end-members need to be defined away from the data cloud. The approach to deal with these non-constant sum data is defined and called Hyperplanar Vector Analysis (HVA). Without abandoning over 50 years of experience, HVA merges the concepts developed over this time and extends the linear unmixing approach to more types of data. The basis for this development involves a translation and rotation of the raw data that conserves information (variability). It will also be shown that HVA is a more appropriate name for both the previous constant sum algorithms and future programs algorithms as well.

W. E. Full (✉)
GXStat, LLC, 1321 Farmstead, Wichita, KS 67208, USA
e-mail: bill@GXStat.com; BillFull@cox.net

© The Author(s) 2018
B. S. Daya Sagar et al. (eds.), *Handbook of Mathematical Geosciences*,
https://doi.org/10.1007/978-3-319-78999-6_42

42.1 Introduction

Unmixing algorithms and programs have been used to solve many different types of geologic problems for more than 50 years. This approach has been developed by geologists for geologists and has been recently ‘borrowed’ by professionals in other fields. For the most part, the International Association for Mathematical Geosciences’ publications *Journal of Mathematical Geology* (later renamed *Journal of Mathematical Geosciences*) and *Computers & Geosciences* have been the venue for the papers describing the developments and computer codes associated with the approaches described in this report. The history of linear unmixing tied to these papers is the topic of this manuscript along with extending the mathematics to make this approach more appropriate for more common types of geologic and petroleum data. The most recent name for these algorithms is Hyperplanar Vector Analysis (HVA)—a name that will be shown to be more appropriate than the other algorithms/program names that have been used in the past.

42.2 History of Constant Sum HVA

42.2.1 *Determination of the Number of End-Members*

The rudiments of HVA started with a report to the Office of Naval Research by Imbrie (1963). In this report, the application of the cosine-theta similarity matrix was defined for the Q-mode factor analysis portions of HVA that were to follow. The cosine is used as a similarity index between two samples (Fig. 42.1a). When the angle between two samples approaches 0.0 (cosine approaches 1.0), the ratio of the two variables are assumed to be nearly the same. Conversely, when a cosine approaches 0.0 ($\Theta = \pi/2$ radians), the two samples are considered very different from each other. In statistics, a cosine value of 0.0 would consider the two samples to be independent of each other. While the Imbrie (1963) approach never calculated a cosine function, it did accomplish the same thing by working with the unit vectors of each sample and with the unit sphere defined by these vectors which was subsequently rotated via an eigenvector rotation. The resulting matrix is the cosine-theta matrix defined for all the samples. Figure 42.1b shows the case where two vectors of differing length would produce a cosine Θ that would indicate that the two vectors would be the same as two vectors of exactly the same length. The constant sum approach assumes that the raw data represents vectors of equal length.

Working with vectors on the unit sphere is one of the fundamental differences between what we have been calling vector analysis and traditional factor analysis. Figure 42.2a illustrates the concept of a unit vector while Fig. 42.2b shows a cross-section of the unit sphere in two dimensions. In traditional factor analysis, in simplified terms, before the eigenvector rotation is performed, the mean of either the raw data or transformed data (usually the z-transform) is subtracted from the

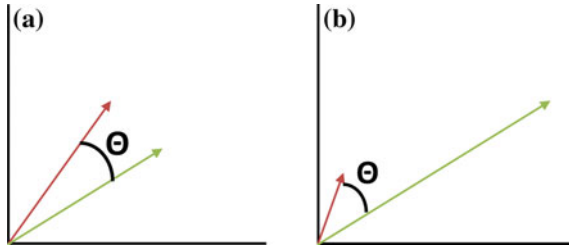


Fig. 42.1 Example of the cosine as a measure of similarity where two samples are very similar to each other in terms of the ratio of the defining variables (a), and where the two samples are more dissimilar than the previous two samples (b). With constant sum models, both set of vectors would be considered as essentially the same

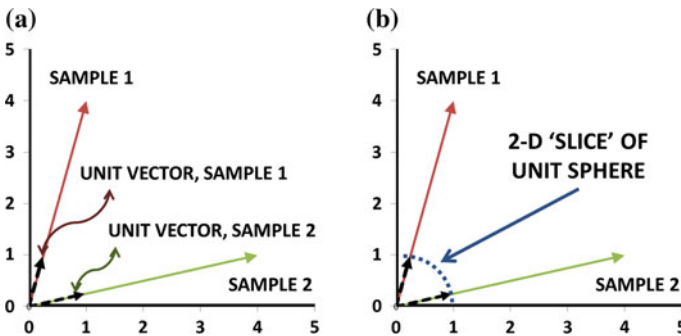


Fig. 42.2 Every sample (row of data) can be considered a vector. The unit vector is the direction of this vector where the length of the unit vector is exactly 1.0 (a). The collections of the sample unit vectors are located on the unit sphere whose radius is 1.0 (b)

variance (or covariance matrix). This step in the procedure is a translation of the axes defining the system (Fig. 42.3). Figure 42.3 also shows in 2-dimensions that the use of the cosine-theta similarity approach does ultimately define eigenvectors and eigenvalues relative to the center of the unit sphere. It should be pointed out that using the approach of Imbrie (1963), the total variability (sum of squares of each coordinate in the space defined by the unit sphere) before and after the eigenvector rotation is simply the number of samples (N). If we have 45 samples, we will have variability in the unit sphere of 45.0. A FORTRAN-IV computer program to perform this procedure was published by Klován and Imbrie (1971) and was named CABFAC (Columbia and Brown Factor Analysis). Unfortunately for a generation of students and practitioners, the terminology used in this and several of the subsequent programs was rooted in factor analysis.

The next step in the evolution of HVA was taken by Miesch (1976a, b). Miesch realized that the CABFAC program was really a combination of linear algebra and geometry. The eigenvector rotation defined by the previous authors was actually

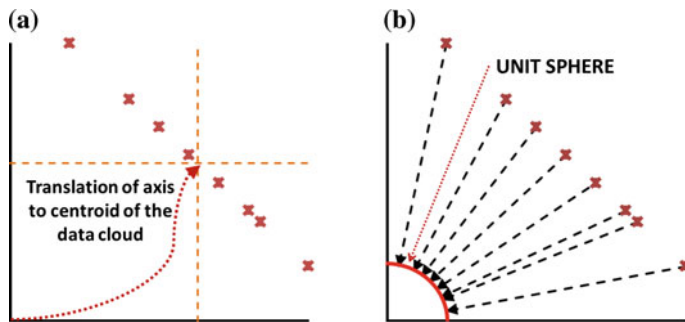


Fig. 42.3 In traditional PCA or factor analysis, the subtraction of the mean is performed before the eigenvector rotation and is a translation of the axes to the center of the data (a). Of course, in a standard PCA or factor analysis, we would divide each value by the standard deviation of the corresponding variable. In contrast, the Q-Mode analysis defined by Imbrie (1963) defines the center of the unit sphere as the point of reference for the eigenvalue rotation (b)

capturing the geometry of the data on the unit sphere. This fact, in conjunction with the observation that with constant sum data the raw samples must fall on either a line (2-D), plane (3-D) or hyperplane (n -D), was a fundamental concept for Miesch. This was a different viewpoint about constant sum data than that reported by Chayes (1971). Miesch concluded that CABFAC can be used to tell us the real dimensionality of the data (must be less than or equal to the number of variables) and that with some additional programming, the end-members and relationships between these end-members and each sample (proportions) can be defined. Programs were created and published by Klován and Miesch (1976) called EXTENDED CABFAC and QMODEL. These two programs, while still using the standard terminology of factor analysis, represented the foundation of the vector analysis unmixing approach that is used to this day. As a matter of fact, rotation procedures such as the orthogonal VARIMAX rotation (Kaiser 1958) are still performed in the programs.

Before we continue with the QMODEL evolution, a discussion of the ways that EXTENDED CABFAC helps us determine the number of appropriate dimensions to choose which is, in reality, the number of end-members present in the data. CABFAC presents us with several ways of defining the exact number or range of end-members that may be present in the data. Note that CABFAC does not tell us anything about what they look like—or the proportions relating these end-members to each sample. For the sake of this discussion, a data set was created wherein four end-members were mixed in known proportions. While the end-members were not constant sum (the sum of each end-member was not the same value), the collection of these data can still be informative, especially when we discuss non-constant sum analysis. The four end-members were taken from NURE stream sediment geochemical samples (Smith 1997) and this data set. For this section on constant sum algorithms, each sample in the data was transformed to a constant value of 1.0 before being submitted to CABFAC/SAWVEC/VECTOR/PVA routines.

The traditional approach used in the past is the scree plot (Fig. 42.4a). In this plot, the user looks for a break in the slope and then interprets this point as the maximum number of end-members present in the data. Note that like real data, Fig. 42.4a shows a case where the scree plot need not behave in an ideal sense. Miesch (1976a, b) recognized that since we are looking at how well the constant sum plane or hyperplane ‘fits’ the original data, back-calculated values from a **reduced** space defined by fewer than n eigenvectors can be directly compared to the variables defined in the raw data or **real** space. This back-calculation simply reverses the mathematics using a reduced number of eigenvectors ‘back’ into the raw data metric via matrix algebra. The comparison is made via the coefficient of determination (CD) function (Draper and Smith 2014) and the CD for each back-calculated variable to the original raw data for a given number of retained eigenvectors is plotted (Fig. 42.4b). Similarly, for each sample, total amount of original variability retained for a given number of eigenvectors is also calculated. This ratio is called the communality for a given sample and is the amount of variability retained by the reduced space divided by the total variability represented by that sample in real space. Figure 42.4c presents a few communality trends for arbitrary samples picked from the test data set. The collection of communalities for a given number of retained eigenvectors can be scanned to look for anomalies that may represent problematic data or the collection can be binned and plotted to assess the range of problems. In the past, a general ‘rule of thumb’ was that, scanning the columns of orthogonal coordinates (loadings) from the fewest to the highest number of end-members, the first time that approximately 5% or less of the data had communalities less than 0.99 and the coordinates had values less than 0.5, then that number of end-members was near the upper range for the maximum number of end-members. The reality was that lower communalities might be due to noise, measurement error, recording error, or it might be the hint of an additional end-member(s) which generally meant it could be more difficult for the modeling programs to define. Johnson (1997a, b) used the insight that by looking at plots of the back-calculated variables to the raw variables, further insights can be gleaned especially by those that want to visualize the ‘pile’ of numbers described earlier. Figure 42.4d displays some of those plots for a single variable. These plots have been called Johnson plots in the programs described later in this report.

Finally, if the assumption is that what is not included is in fact noise, there might not be enough information available that can be used to define any additional end-members. In such a case, the distribution of the variability relative to each ‘removed’ eigenvector can be examined. This is usually done by looking either at the ‘coordinates’ of the removed eigenvectors (similar to looking at the principal component loadings in Principal Components Analysis) and using external tools such as JMP Pro (1989–2017). The latest programs create appropriate data tables for this step, and for all of the previous steps with key information, that can be used in ancillary programs that have many more statistical functions and better graphics. One such example might be to examine the behavior of the ‘removed’ eigenvector coordinates to verify that the ‘removed’ eigenvectors do not contain meaningful information (i.e. whether they can be considered noise and not pertinent to the

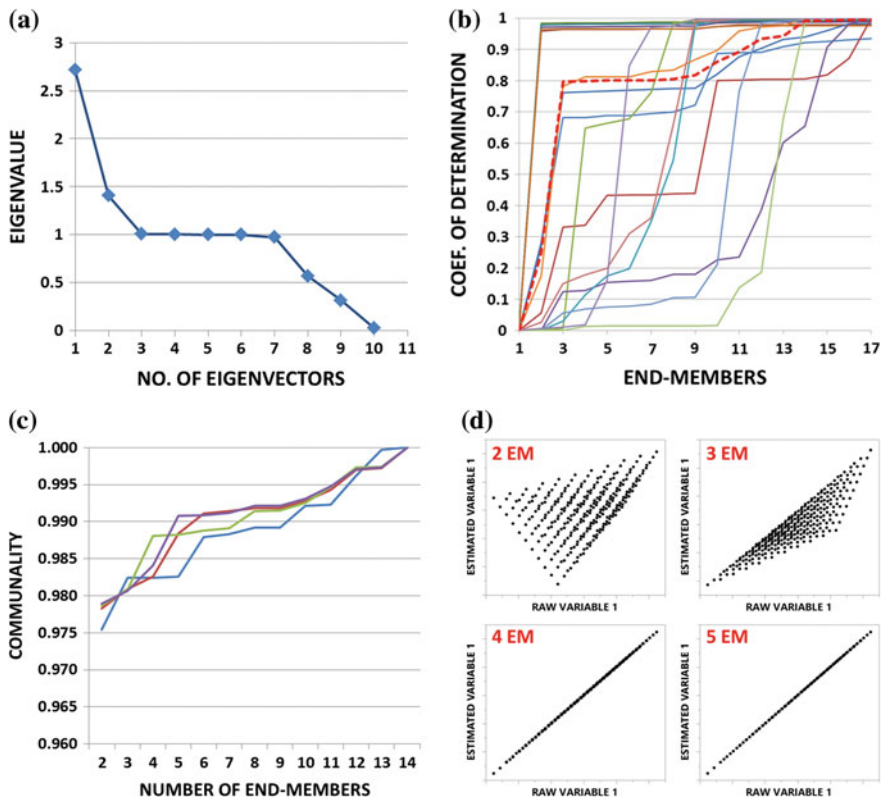


Fig. 42.4 An example of the scree plot from the test data where the number of eigenvectors retained are plotted against individual eigenvalues (a). A plot of the CD's for the test data shows how each variable contributes to the overall choice of the number of end-members (b). Communalities for four samples are presented for the range of eigenvectors retained (c). Collection of Johnson plots showing the visual fit relative to a single variable as the number of end-members (EM) has increased (d)

overall model). The user would have to a priori establish criteria that defined noise in terms of the individual data used and/or by some distribution parameters such defined by mean and standard deviation, for instance.

42.2.2 Determination of the Composition of the End-Members and Proportions

Klovan and Miesch (1976) developed the program QMODEL based on Miesch (1976a) in order to define the composition of the end-members and calculate the proportions relating each individual sample to this set of end-members. Given the

choice of the number of end-members normally based on EXTENDED CABFAC, the procedure to define the compositions and proportions (oblique coordinates of the space defined by the end-member axes) is strictly linear algebra. The mathematics used up to this point is well defined in Miesch (1976a). QMODEL was designed to be a data modeling program that required interaction with the user. A discussion of these approaches and other alternatives can be found in Clarke (1978). There were several ways for this program to define end-members:

- (1) Use the retained eigenvectors as end-members (principal factors)
- (2) Use the VARIMAX axes as the end-members (VARIMAX factors)
- (3) Use Imbrie's oblique end-members (the extremes in the reduced space—EXNORC routine)
- (4) Use the extremes as defined by the back-calculated extremes in the raw space—the EXRAWC routine
- (5) Define the end-members by the row indices of the set of samples (e.g. use the 5th and 12th sample as end-members)
- (6) Define the actual composition of each of the end-members (these would normally be a set of end-members defined in the raw metric that the user would want to test)
- (7) Externally define the end-members by defining the VARIMAX coordinates (loadings)—this would normally be done when the user has made multiple plots of the data in VARIMAX space

For each of the choices in the original QMODEL program, correct choices produced end-members that were realistic (defined by acceptable variables in the raw data space) and by proportions that were between 0.0 and 1.0. Problems arose with many data sets when the raw end-member compositions were unrealistic and/or the proportions were out of range. This problem is commonly encountered when there are many variables and samples which makes visualization of the location of the potential end-members difficult at best. To that end, new modeling approaches were devised that gave some automation toward the definition of proper end-members and proportions.

Full et al. (1981, 1982) devised two alternative methods that involves an iterative scheme that started with one of the original QMODEL choices above or with fuzzy cluster centers (Bezdek et al. 1984), and then allowed the program to define end-members external to the data, check their proportions for viability, change if needed the set of end-member compositions to the nearest viable location, and repeat the process until either the program shows no convergence or an acceptable solution is reached. The goal was to determine appropriate sets of end-members closest to the data cloud defined by the samples. This may be likened to trying to minimize the area or hyper-area that represents the planar/hyperplanar convex hull defined by the end-members. The computer code, along with some bug fixes to the EXRAWC and EXNORC subroutines, can be found in the appendix of Full (1981). A general discussion of these methods and their applications at the time can be found in Ehrlich and Full (1988). Alternatives to the aforementioned approaches

can be found in Leinen and Pias (1984) and Weltje (1997). Insights into the appropriate applications of these algorithms and recognizing how to detect problems with the underlying model were discussed in Williams et al. (1988a, b, Chaps. 15 and 19). Optimized data binning for continuous distributions that improved the results of these algorithms were presented in Full et al. (1984).

42.2.3 The Renaming to Polytopic Vector Analysis

In the early 1980s, given the changes to the original CABFAC and QMODEL programs, the approach was renamed SAWVEC (South Carolina and Wichita Vector Analysis) and sometimes simply VECTOR. It was the recognition that the algorithms were dominated by vector algebra that prompted the name change. Circa 1990, the exact same approach was further renamed Polytopic Vector Analysis and applied under that name in Evans et al. (1992) and in many of the references mentioned in later in this report. Around this time, Sterling James Crabtree, then at the University of South Carolina, translated the FORTRAN IV code of Full (1981) into the C programming language and developed a Windows interface and ultimately called the program PVA. This program can be recognized by the fact that the first step after starting the program was to resize the introductory window.

The use of the term polytope has been problematic for this author even though the term was used in the original Full et al. (1981) algorithm. The field of polytopic mathematics has been around for over a century and was generally formalized by Coxeter (1948, 1973). Coxeter assumed that a polytope was a geometric construct in 4 or more dimensions with the degenerate cases being the point in 0 dimensions, the line segment in 1 dimension, the polygon in 2 dimensions and polyhedron in 3-dimensions representing polytopes of dimension 0, 1, 2 and 3 respectively. A search of the literature on polytopes shows that this field of mathematics is rich in various definitions of a polytope, depending for instance on whether you are talking about a convex hull in n -dimensions or more complex surfaces as in star-type polytopes. It is clear that for the geologist this can be a confusing landscape to travel through. A simplistic definition would be that a polytope is an n -dimensional geometric figure ($n > 3$) whose sides are planes or hyperplanes. The implicit assumption is that a polytope has some kind of volume or hypervolume. Henk et al. (1997) even developed equations for calculating this volume or hypervolume for many types of regular polytopes.

If a polytope can be considered as a region of n -dimensional space that is enclosed by hyperplanes (Coxeter 1973), then that causes problems for linear unmixing. If we consider a vector emanating from a point outside that region and look at the potential intersections of that vector with the polytope, the only possibilities for unique points would be if the vector intersected the vertices of the polytope. If the vector intersected a side, there could possibly be two or more points of intersections which would cause havoc with the uniqueness aspects of the unmixing model. The reality is that in the non-constant sum model, regardless of

the number of dimensions (end-members), the data fall on a hyperplane when the number of dimensions is greater than 3. As we will see later, it is this fact that the extension of all of the previous algorithms to non-constant sum data can be realized. Because of the confusion associated with the term ‘polytope’ relative to the understanding of the previously described algorithms, they have been renamed Hyperplanar Vector Analysis (HVA).

42.2.4 Review of the Applications of Constant Sum Unmixing

The CABFAC, EXTENDED CABFAC-EXTENDED QMODEL, SAWVEC, VECTOR, PVA algorithms and programs (henceforth referred to as HVA family of algorithms) have found application in many geologic disciplines. Some of the earliest studies have involved the analysis of size data in both nearshore and lacustrine environments. These include the work of Klován (1966) and Solohub and Klován (1970) using traditional sieved size data. Fillon and Full (1984) used specialized equipment to define the size of particles on an individual basis and defined 5 different sources of deep sea sediment. As pointed out in Fillon and Full (1984) and Full et al. (1984), the success or failure of size analysis depends on the optimization of the size data using transforms such as the maximum entropy method.

In the field of grain shape analysis, the heart of the analytic scheme was the constant sum unmixing algorithms described above. The studies included sediment from Monterey Bay, CA (Porter et al. 1979). Brown et al. (1980), Reister et al. (1982), Mazzullo et al. (1982, 1984), Hudson and Ehrlich (1980), Smith et al. (1985), Tortora et al. (1986) and Evangelista et al. (1986, 1994, 1996) looked at sediment distributions along beaches, barrier islands, shelf and abyssal plains. Murillo-Jiménez et al. (2007) examined the sediment from a relatively large region along the southern coast of Baha California, MX. Material from more lithified material was studied by Mazzullo and Ehrlich (1980, 1983) and Civitelli et al. (1992). El-Awawdeh and Full (1996) looked at changes in key morphology in Florida Bay over time. The methods used in those studies were reviewed in Ehrlich and Full (1984a, b) and Zhao et al. (2004).

The biologic morphology and fossil assemblage scientists were early adapters of the HVA family of algorithms. Healy-Williams (1983, 1984) and Healy-Williams et al. (1997) worked with forams, Burke et al. (1986) with ostracodes and Kensington and Full (1994) with scallops. Williams et al. (1988a, b) looked at correlations of foram shapes with isotopic signatures. Assemblages of microfossils were unmixed in Gary et al. (2005) and Zellers and Gary (2007).

A major area of investigation using the HVA family of algorithms deals with environmental science. Detecting contaminants in soils and identifying their sources was reported by Ehrlich et al. (1994), Wenning and Erickson (1994), Doré et al.

(1996), Jarman et al. (1997), Johnson (1997a, b), Huntley et al. (1998), Bright et al. (1999), Johnson et al. (2000, 2001), Johnson and Quensen (2000), Nash and Johnson (2002), Nash et al. (2004), Barabas et al. (2004a, b), Magar et al. 2005, DeCaprio et al. (2005), Towey et al. (2012), Leather et al. (2012) and Megson et al. (2014). The Battelle Memorial Institute (2012) has listed PVA in their handbook for determining the sources of PCB in sediments.

The HVA family of algorithms is critical for the field of PIA (Petrographic Image Analysis). The literature includes Ehrlich and Horkowitz (1984), Ehrlich et al. (1984, 1991a, b, 1996, 1997), Ross et al. (1986), Scheffe and Full (1986), Full (1987), Etris et al. (1988), McCreesh et al. (1991), Ross and Ehrlich (1991), Fern et al. (1993), Bowers et al. (1994, 1995), James (1995), Carr et al. (1996), Yannick et al. (1996), Anguy et al. (1999, 2002) and Sophie et al. (1999).

Igneous rock researchers have also been an adapter of these unmixing algorithms. These include Horkowitz et al. (1989), Statterger and Morton (1992), Tefend et al. (2007), Vogel et al. (2008), Deering et al. (2008), Barclay et al. (2010), Szymanski et al. (2013), Lisowiec et al. (2015) and most recently by Blum-Oeste and Wörner (2016).

The unmixing of sources of oil using the HVA algorithms has been reported by Collister et al. (2004), Van de Wetering et al. (2015), Abrams et al. (2016) and Mudge (2016). The correlation between stratigraphy and chemical stratigraphic data was explored by McKenna et al. (1988). “Quasigeostopic potential vorticity” was explored in Evans et al. (1992). Mason and Ehrlich (1995) looked at aspects of well logs for basin exploration (1995). Full and James (2015) used the HVA (non-constant sum version) to decompose a large data set consisting of exploration data in order to better assess exploration and exploitation risk. At least two patents have mentioned using the HVA family of algorithms for analysis of the data derived from their process (Shafer and Ehrlich 1986; Nelson et al. 2013).

The above literature is by no-means the entire community of users of the unmixing approach began by Imbrie (1963). There have been verbal reports of researchers doing work with Shakespeare’s plays, classifying business reports, analyzing social data and even applying these approached to marketing data. The success or failure of these studies cannot be directly ascertained, but represent some interesting applications.

42.3 Non-constant Sum Data and Algorithms

The previous sections, for the most part, dealt with rows of data whose row sum was the same or very similar for each sample (vector). This type of data is merely a subset of the data commonly encountered in the geologic sciences and, if you want to use the previous algorithms, you have to potentially degrade your data by transforming it to percentages or some other appropriate singular value. Oftentimes, this involves removing the absolute quantity involved with each sample. For example, if you have six glasses and pour into each glass a variable amount of three

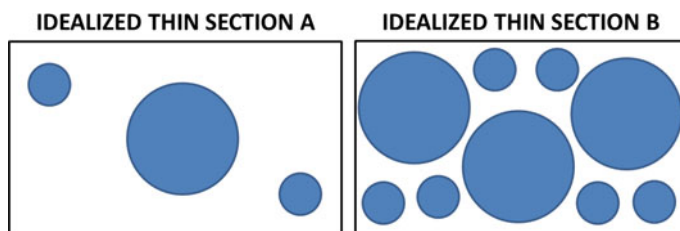


Fig. 42.5 An example of two idealized images that would produce the same smooth-rough distributions in the petrographic image analysis system described in Ehrlich (1991a, b). Note that in image **a**, the porosity would be much greater than image **b** which would greatly affect the calculation of permeability and other petrophysical variables

solutions, some glasses might contain a greater volume and some a lesser volume—here the quantity of each solution might be important. The concept of unmixing might still be appropriate but would only be accurately defined in terms of end-member compositions and sample proportions in very special cases that will be discussed below. With petrographic image analysis which heavily uses the unmixing algorithms, two collections of imaged thin sections with vastly different porosities would ultimately have equal constant sum smooth-rough distributions (Fig. 42.5). Petrophysical logs, formation depths, seismic parameters and other petroleum related data are mostly non-constant sum in nature. There are many other types of data where the concept of mixtures and unmixing can be validly applied.

What happens when you try to apply the constant sum programs to inherently non-constant sum data? This topic was partially addressed by Klován (1981) without addressing the application of determining end-members and proportions using the techniques described by Full et al. (1981, 1984). In his paper, Klován notes that, if the data can be approximated by a plane or hyperplane parallel to the constant sum plane, then the aforementioned algorithms can be appropriately applied. However, Klován (1981) acknowledges problems when the surface defined by the non-constant sum data is not parallel to the unit constant sum plane. Some of the problems can be demonstrated by a simple diagram in two dimensions (Fig. 42.6). Note that the midpoint of the non-constant sum segment does not correspond to the midpoint of the constant sum plane which would be the proportions reported for this point by the computer codes. Using some of the usual functions to create constant sum data that are available in the program would not help matters. A more complex series of transformations using trigonometry could be easily developed for 2 or 3 dimensions but would be difficult to visualize and cannot be easily generalized to n dimensions. Also note that Fig. 42.6 represents an example in two dimensions which intersects the two axes making the determination of end-member compositions a bit easier; they would be represented by the end-points of each line and whose compositions would be the raw data points defining these end-points. If end-members needed to be defined beyond the data

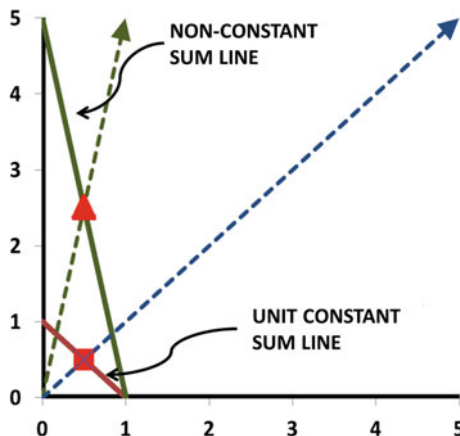


Fig. 42.6 A simplistic example of some of the issues associated with using constant sum algorithms with non-constant sum data. The unit constant sum line is represented by the solid line passing through the points (1, 0) and (0, 1). The non-constant sum data is represented by the solid line at an oblique angle to the constant sum plane. The mid-points (0.5, 0.5) proportion of each line is represented with a symbol. Note that the extended unit vector (represented by the dashed line) that represents the midpoint of the constant sum system is divergent from the same unit vector that passes through the mid-point of the non-constant sum line segment

cloud, the definition of the end-member compositions would be very difficult when there are more than 3 dimensions.

How to deal with the non-constant sum problem was solved in the mid-1980s and has been used in petroleum industry projects and for research projects for the Department of Defense. The code was initially run on a 386-processor with 387-co-processor as well as IBM mainframes. It is only recently that the computer code has been written for Windows operating system with a Windows GUI. The abstract concept behind the approach to dealing with this type of data is to recognize that ultimately any mixing problem deals with data on either a line segment (in 2-d), a plane (2 or 3-d) or hyperplane in more than 3 dimensions. The goal then is to define that hyperplane and translate/rotate the data to a plane/hyperplane that is parallel to the unit constant sum plane where we can apply the usual constant sum approaches. Afterward, any time we want to know what the raw compositions are, we reverse the translation/rotation to bring us back into the original metric. In this way, the earlier approaches are not abandoned but can be efficiently extended to almost any other data that can be modeled as a mixture.

The procedure for this translation/rotation is the following:

- (1) Remove the mean from the data. This is equivalent to the first step of principal components (Davis 2002; Draper and Smith 2014). The visualization for this step is that the axes defining the raw data are translated to the mean of the data with no loss of information.

- (2) An eigenvector rotation is performed on this data. If we were to divide the variable standard deviation by each corresponding row of the matrix defined from the previous step above before this eigenvector rotation, we would have a standard principal component analysis. Since we have not done so, we have not altered the absolute position of the raw data in the data cloud nor the variance associated with the raw data—no loss of information. It should be noted that this step of the analysis is performed by the SVD computer algorithm (Golub and Reinsch 1970) programmed to use quad precision (128 bit) to minimize any information loss and to be able to run large raw data matrices. The rest of the HVA program currently runs in double precision.
- (3) Create a new matrix G with the following definition:
Letting $ANV = 1/NV$ where NV is the number of variables and $ANX = \text{SQRT}(1 - ANV)$, then G is defined as an $NV \times NV$ matrix with every element $-ANV/ANX$ except along the main diagonal where the element is $(1 - ANV)/ANX$. Note that the sum of squares of each row element is 1 and each of the elements is orthogonal and represents spanning vectors for the constant sum plane.
- (4) Using the Gram-Smith orthogonalization procedure (Cheney and Kincaid 2008), orthogonalize the matrix defined in the previous step. Call this matrix G^0 .
- (5) Create a new matrix G^* where $G^* = G^0 * B$ where B is the set of previously defined eigenvectors in step 2. Note that since G^* is an orthogonal matrix, then $G^{*-1} = G^{*T}$ where T is the notation for transpose (this fact is well known in mathematics: see for example Schwartz 2011). G^* and G^{*T} gives us the mechanism to go from the raw data space to a plane parallel to a constant sum plane. However, since this new reference system also contains the origin, the addition of a constant value will translate the plane/hyperplane away from this origin by a constant value to a position parallel to the constant sum plane/hyperplane. In the program, this constant value is called $AVAR$ and, based on experience, has been set to $2 * NV *$ (smallest value of the G^* rotated coordinates) or 1.0 if this number is lower than 1.0.

In more simplistic terms, what we have done is to create an $NV \times NV$ matrix ($NV =$ the number of variables) that will be used to rotate the raw data in order to create a one-to-one correspondence with a set of points in a plane/hyperplane parallel to a constant sum plane/hyperplane. This matrix was orthogonalized and the application of this rotation and translation results in the loss of no information. Since this is an orthogonal matrix, the transpose of this matrix is the inverse of the matrix and gives us the function to go from the constant sum hyperplane to the raw data. These functions allow for properly defined proportions and end-member compositions whether the end-members are contained in the data or not. Figure 42.7 illustrates what the procedure is doing in general.

The constant sum routines can then be applied as they were before only using the G^* and G^{*T} matrix defined above to move from the raw data hyperplane to the constant sum hyperplane with no (or minimum loss due to computational error) loss of information. This approach capitalizes on more than a half-century of previous

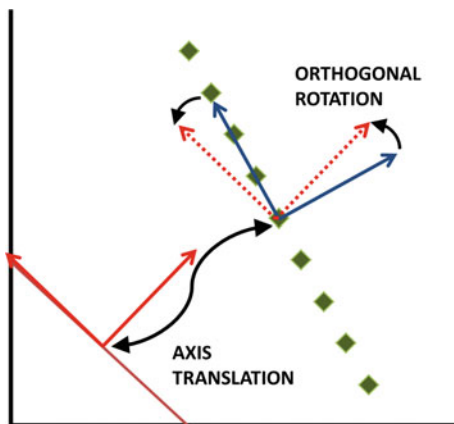


Fig. 42.7 A 2-dimensional representation of the procedure to define the G^* matrix procedure described in the text. Note that in 2-dimensions, the first eigenvector defines the direction of the line segment and the second the normal to this segment. The red axes represent the first eigenvector and the normal to the constant sum line. These axes are then translated to the mean of the non-constant sum data cloud defined by the green diamonds. The blue axes represent the first eigenvector and the normal to the non-constant sum line. This set of axes will be orthogonally rotated to the position of the constant sum axes (dotted axes), (i.e., the raw data will be defined by a new set of coordinates). Mathematically, this procedure will not result in information loss

algorithmic and programming experience. Furthermore, the appropriateness of the unmixing model in non-constant sum space can be checked by looking at the set of eigenvalues—data that do not fall on the mixing hyperplane will have a value other than 0.0 for the last eigenvalue. Additionally, by checking the raw data on a sample-to-sample basis with its equivalent location in the constant sum hyperplane via a similar function to the communality will allow the user to examine potentially aberrant data.

As a demonstration sample, using the previously defined test data set, we can compare the end-members and proportions when they are subjected to a constant sum approach (data was transformed to 100%) and the non-constant sum approach. The set of end-members are shown in Table 42.1 and randomly selected proportions for 10 of the original 296 samples are tabulated in Table 42.2. This data set will be made available from the GXStat website (www.GXStat.com). Note that these data contained the end-members as samples and therefore no iterative schemes such as those described in Full et al. (1981, 1984) were used. It should be noted that, for the most part, the end-members are not that extreme compared to potential test end-members that could have been chosen. Mathematically, this is saying that, with the test data used in this example, most of the variables in the mixing hyperplane lie in portions of that hyperplane which can be modeled as constant sum (i.e. take away the handful of variables that lie in a section of the hyperplane that is most oblique to the constant sum plane, and the data might be able to be modeled using the constant sum algorithm). In the more common case

Table 42.1 Test data set end-members (TEST EM) with constant sum end-members (CS EM) and the HVA non-constant sum end-members (HVA EM). Note the disparity between the constant sum end-members (central gray area) and the actual model end-members (white area on right of table). Also note how close these actual model end-members and the HVA end-members reflect each other. The variables represent parts per million data of 20 different elements

VAR. #	TEST EM 1	TEST EM 2	TEST EM 3	TEST EM 4	CS EM 1	CS EM 2	CS EM 3	CS EM 4	HVA EM 1	HVA EM 2	HVA EM 3	HVA EM 4
1	184.0	171.0	204.0	197.0	17.1	27.8	28.8	26.1	184.0	171.0	204.0	197.0
2	0.6	0.3	0.5	0.5	0.1	0.1	0.1	0.0	0.6	0.3	0.5	0.5
3	1.9	6.4	11.0	4.0	0.3	0.1	1.9	1.0	1.9	6.4	11.0	4.0
4	710.0	320.0	280.0	340.0	66.1	47.5	39.6	49.0	710.2	320.5	280.2	340.1
5	4.0	5.0	2.0	16.0	0.2	3.5	-0.5	0.6	4.0	5.0	2.0	16.0
6	29.0	29.0	32.0	27.0	2.7	3.3	4.6	4.8	29.0	29.0	32.0	27.0
7	2.2	9.0	8.0	5.8	0.2	0.6	1.0	1.8	2.2	9.0	8.0	5.8
8	17.0	22.0	35.0	26.0	1.7	3.3	5.4	3.0	17.0	22.0	35.0	26.0
9	1.4	0.7	1.9	1.0	0.2	0.1	0.3	0.0	1.4	0.7	1.9	1.0
10	3.0	7.0	14.0	8.0	0.4	0.9	2.3	0.9	3.0	7.0	14.0	8.0
11	0.5	0.4	0.6	0.5	0.0	0.1	0.1	0.1	0.5	0.4	0.6	0.5
12	6.1	3.0	6.2	7.0	0.6	1.2	0.9	0.2	6.1	3.0	6.2	7.0
13	9.0	2.6	2.8	3.7	0.9	0.6	0.4	0.3	9.0	2.6	2.8	3.7
14	0.1	0.1	0.0	0.0	0.0	0.0	0.0	0.0	0.1	0.1	0.0	0.0
15	11.0	12.0	19.0	14.0	1.1	1.8	3.0	1.6	11.0	12.0	19.0	14.0
16	0.2	0.2	0.2	0.1	0.0	0.0	0.0	0.0	0.2	0.2	0.2	0.1
17	4.0	12.0	11.0	11.0	0.3	1.5	1.3	2.2	4.0	12.0	11.0	11.0
18	4.0	24.0	14.0	12.0	0.1	1.1	1.4	5.3	4.0	24.0	14.0	12.0
19	80.0	30.0	50.0	45.0	7.8	6.2	8.3	2.8	80.0	30.0	50.0	45.0
20	2.8	2.6	6.2	4.0	0.3	0.5	1.0	0.2	2.8	2.6	6.2	4.0

Table 42.2 Ten randomly selected samples were picked to show the trends of the proportions from results of the application of the constant sum algorithms and the non-constant sum programs. The a prior proportions for each end-member used to create each test sample is given by the columns ORIG. PROP., the constant sum derived proportions by the grayed columns labeled CS PROP in the center of the table and the results of the non-constant sum application is given by the columns on the right of the table labeled HVA PROP. The average error for the proportions of the HVA results was found to be ± 0.0004 which was largely attributed to the fact that the raw data was defined using two decimal-point accuracy. The sample numbers represent the sequence number of the row of the test data set

SAMP. #	ORIG. PROP. EM 1	ORIG. PROP. EM 2	ORIG. PROP. EM 3	ORIG. PROP. EM 4	CS PROP. EM 1	CS PROP. EM 2	CS PROP. EM 3	CS PROP. EM 4	HVA PROP. EM 1	HVA PROP. EM 2	HVA PROP. EM 3	HVA PROP. EM 4
14	0.70	0.00	0.00	0.30	0.78	0.00	0.00	0.22	0.70	0.00	0.00	0.30
32	0.60	0.10	0.20	0.10	0.70	0.07	0.15	0.08	0.60	0.10	0.20	0.10
81	0.40	0.10	0.40	0.10	0.51	0.08	0.33	0.08	0.40	0.10	0.40	0.10
95	0.30	0.60	0.10	0.00	0.41	0.50	0.09	0.00	0.30	0.60	0.10	0.00
107	0.30	0.10	0.30	0.30	0.40	0.08	0.26	0.26	0.30	0.10	0.30	0.30
121	0.20	0.40	0.30	0.10	0.28	0.34	0.28	0.09	0.20	0.40	0.30	0.10
154	0.20	0.20	0.00	0.60	0.28	0.17	0.00	0.55	0.20	0.20	0.00	0.60
166	0.20	0.10	0.40	0.30	0.28	0.08	0.36	0.28	0.20	0.10	0.40	0.30
208	0.10	0.20	0.00	0.70	0.15	0.18	0.67	0.00	0.10	0.20	0.00	0.70
252	0.00	0.50	0.30	0.20	0.00	0.48	0.31	0.21	0.00	0.50	0.30	0.20

where end-members need be defined external to the data cloud, the results would have been potentially far off and confused if the constant sum algorithm was applied. Also note that if the user did use the constant sum routines to define the composition of the end-members and either manually extracted the raw data of an internal end-member or the 'nearest' actual point (defined by the raw data) to the external end-member, it would be difficult to know how these points relate to all of the other data samples—the user would simply not know if all the data truly fall on a mixing plane or hyperplane. Finally, because HVA rotates the data to a plane parallel to the constant sum plane, when the data are inherently constant sum, no new program is needed.

Finally, it should be noted that this non-constant sum model will work for any mixing system that can be modeled as a plane or hyperplane. The dimensionality of the hyperplane must be less or equal to than the number of variables otherwise there will not be a unique solution to the end-member and proportions problem. This does bring up the case where a three end-member solution (defined by a triangle) in two dimensions can be solved using these algorithms. The G^* rotation described above can potentially produce a plane or hyperplane that intersects with the origin defining an end-member consisting of the origin with (0, 0, ...) as its composition. The interpretation of the origin as an end-member has been successful in previous studies when this situation has been encountered. It can be, however, a tricky proposition depending on the type of data being analyzed. It might be useful to substitute a value close to the origin for the definition of that end-member instead of using the origin as an end-member composition.

Areas of application of this approach have included chemo-stratigraphic data, correlation and mapping of wireline well logs, unmixing of oil compositions preserving volume of source material, determination of various forms of risk in exploration schema, correlating biologic assemblages to seismic stratigraphy, and determination of 'sweet spot' locations for oil exploitation, to name a few. Unfortunately, the results of these reports remain confidential. It is anticipated that these and new applications will be reported in the future in various literature.

42.4 Summary

Fifty years of research and development have given the geologic community a useful tool for the analysis of mixtures. It is anticipated at this time that this approach will last well into the future, especially since the program will be made available to anyone in any field they want. It should be noted however, that there are still untested areas of research in this field. The most appropriate approach for the definition of extreme end-members is still an open discussion. Generally, researchers have been looking at the extremes of the data and not looking so much at the bulk of the data. While much of the variable density of the raw data may be due to localized over-sampling problems (usually, we geologists sometimes just analyze the data we have!), there are other methods such as FUZZY clustering (Full

et al. 1984; Bezdek et al. 1984) and algorithms that use FUZZY variables to define data density in terms of sets of point, lines, planes, hyperplanes and various n-dimensional spaces (Bezdek 1981).

Another area that needs some additional work is the definition of new criteria that will allow the various iterative schemes to know when the ‘best’ solution is achieved, when there might not be a complete convergence. In terms of computer programming, what would be beneficial is to be able to define one or more ‘fixed’ end-member(s) (the number being less than the original number of chosen end-members) and let the program determine other potentially viable end-members using the DENEG iteration scheme (i.e. one or more end-members want to be fixed in the analysis—the programs have always had ways of externally defining all of the end-members). Additionally, defining how the end-members interact with the modeled environment (such as when a geochemical component reaches a given level and precipitates out of the system) would also be of great use. This has been accomplished in the past by making alterations to the program, recompiling the code and proceeding with the newly built custom program. Being able to run this option without having to recompile would be quite useful. Another item on the wish list would be to convert the program out of FORTRAN IV, although the current program is very fast and FORTRAN has become a versatile programming language. This author acknowledges that there are fewer and fewer people who can program in this language, especially in the Windows environment. A language that has a ‘better’ future would be of great advantage, especially since the programs and algorithms may be used by a wider audience. Additionally, all of the mathematics needs to be described in one place along with a user manual that describes in detail not only all the options but also the whys and wherefores of particular options. It should be noted that the program has a built-in user manual but does not go into details of the more subtle nuances associated with the algorithms. These missing discussions will be the topic of various discussions available on the GXStat website (www.GXStat.com). There is even some progress in producing an R version of the program for those who want to incorporate this approach into their projects. This flexibility will be of benefit to a large community of potential practitioners.

Finally, there is something that can be gleaned from the list of references. The access of researchers to the HVA family of algorithms has been somewhat limited by both changes in the computer industry (computer languages and graphic user’s interfaces in addition to hardware) and by research association (i.e. who you know). It is for this reason that the complete source code and compiled code for the past algorithms and the HVA code discussed in this report will be made freely available from the GXStat website (www.GXStat.com) or directly from the author. This, in addition to the test data set and additional research programs such as FUZZY n-Varieties written by this author, will also be made available (in FORTRAN, of course) through this outlet. This open access will allow others to contribute to the mathematics and algorithms, making them even more useful for the next 50 years.

Acknowledgements Many people have contributed to the development of the HVA family of algorithms and programs. It was one of the intents of this report to give them the due credit. Also credit goes to those that have spent a great deal of their career to the application and dissemination of the unmixing approach. To that end, the grand prize should go to Professor Robert Ehrlich, without whom many would not have known about the diverse applications of this approach. I would like to thank Drs. Magdalena and Nils Blum-Oeste for their comments and improvements on this manuscript and overall support along with pushing the subject material. Dr. Lucinda Brothers-Full also help with editing this manuscript. Finally, I would like to apologize profusely to anyone who felt I purposely left them off the list of references.

References

- Abrams MA, Greb MD, Collister JW, Thompson M (2016) Egypt far Western Desert basins petroleum charge system as defined by oil chemistry and unmixing analysis. *Mar Pet Geol* 77:54–74
- Anguy Y, Belin S, Bernard D, Fritz B, Ferm JB (1999) Modeling physical properties of sandstone reservoirs by blending 2D image analysis data with 3D capillary pressure data. *Phys Chem Earth (A)* 24:581–586
- Anguy Y, Belin S, Ehrlich R, Ahmadi A (2002) Interpretation of mercury injection experiments using a minimum set of porous descriptors derived by quantitative image analysis. *Image Anal Stereol* 21:127–132
- Barabás N, Goovaerts P, Adriaens P (2004a) Modified polytopic vector analysis to identify and quantify a dioxin dechlorination signature in sediments: 1. Theory. *Environ Sci Technol* 38:1813–1820
- Barabás N, Goovaerts P, Adriaens P (2004b) Modified polytopic vector analysis to identify and quantify a dioxin dechlorination signature in sediments. 2. Application to the Passaic River. *Environ Sci Technol* 38:1821–1827
- Barclay J, Herd RA, Edwards BR, Christopher T, Kiddle EJ, Plail M, Donovan A (2010) Caught in the act: implications for the increasing abundance of mafic enclaves during the recent eruptive episodes of the Soufrière Hills Volcano, Montserrat. *Geophys Res Lett* 37:L00E09:1–L00E09:5. <https://doi.org/10.1029/2010gl042509>
- Battelle Memorial Institute et al (2012) A handbook for determining the sources of PCB contamination in sediments: Technical report TR-NAVFAC EXWC-EV-1302, Technical report TR-NAVFAC EXWC-EV-1302, NAVFAC Engineering and Expeditionary Warfare Center, Port Hueneme, CA, 164 pp
- Bezdek JC (1981) Pattern recognition with fuzzy objective function algorithms. In: *Advanced applications in pattern recognition*. Springer Science & Business Media, 272 pp
- Bezdek JC, Ehrlich R, Full WE (1984) FCM: the fuzzy c-means clustering algorithm. *Comput Geosci* 10:191–203
- Blum-Oeste M, Wörner G (2016) Central Andean magmatism can be constrained by three ubiquitous end-members. *Terra Nova* 28:434–440
- Bowers MC, Ehrlich R, Howard JJ, Kenyon WE (1995) Determination of porosity types from NMR data and their relationship to porosity types derived from thin section. *J Pet Sci Eng* 13:1–14
- Bowers MC, Ehrlich R, Clark R (1994) Determination of petrographic factors controlling permeability using petrographic image analysis and core data, Satun Field, Pattani Basin, Gulf of Thailand. *Mar Pet Geol* 11:148–156
- Bright DA, Cretney WJ, MacDonald RW, Ikonomou MG, Grundy SL (1999) Differentiation of polychlorinated dibenzo-*p*-dioxin and dibenzofuran sources in coastal British-Columbia, Canada. *Environ Toxicol Chem* 18:1097–1108

- Brown PJ, Ehrlich R, Colquhoun DJ (1980) Origin of patterns of quartz sand types on the southeastern United States continental shelf and implications on contemporary shelf sedimentation: Fourier grain shape analysis. *J Sediment Res* 50:1095–1100
- Burke CD, Full WE, Gernant RE (1986) Recognition of fossil fresh water ostracodes: Fourier shape analysis. *Lethaia* 20:307–314
- Carr MB, Ehrlich R, Bowers MC, Howard JJ (1996) Correlation of porosity types derived from NMR data and thin section image analysis in a carbonate reservoir. *J Petrol Sci Eng* 14:115–131
- Chayes F (1971) Ratio correlation: a manual for students of petrology and geochemistry, Chicago and London. University of Chicago Press, 99 pp
- Cheney W, Kincaid DR (2008) Linear algebra: theory and applications, 1st edn. Jones & Bartlett Learning, 740 pp
- Civitelli G, Corda L, Evangelista S, Full WE (1992) Fourier quartz shape analysis: application to terrigenous sediments of Laga and Cellino Formations (Central Italy). *Bollettino della Societa Geologica Italiana* 111:355–366
- Clarke TL (1978) An oblique factor analysis solution for the analysis of mixture. *Math Geol* 10:225–241
- Collister J, Ehrlich R, Mango F, Johnson GW (2004) Modification of the petroleum system concept: origins of alkanes and isoprenoids in crude oils. *Am Assoc Pet Geol Bull* 88:587–611
- Coxeter HSM (1948) Regular polytopes, London, Methuen, 341 pp
- Coxeter HSM (1973) Regular polytopes, 3rd edn. Dover Publications, 368 pp
- Davis JC (2002) Statistics and data analysis in geology, 3rd edn. Wiley, 638 pp
- DeCaprio AP, Johnson GW, Tarbell AM, Carpenter DO, Chiarenzelli JR, Morse GS, Santiago-Rivera AL, Schymura MJ, Akwesasne Task Force on the Environment (2005) PCB exposure assessment by multivariate statistical analysis of serum congener profiles in an adult Native American population. *Environ Res* 98:284–302
- Deering CD, Cole JW, Vogel TA (2008) A rhyolite compositional continuum governed by lower crustal source conditions in the Taupo Volcanic Zone, New Zealand. *J Petrol* 49:2245–2276
- Doré TJ, Bailey AM, McCoy JW, Johnson GW (1996) An examination of organic/carbonate-bound metals in bottom sediments of Bayou Trepagnier, Louisiana. *Trans Gulf Coast Assoc Geol Soc* 46:109–116
- Draper NR, Smith H (2014) Applied regression analysis. Wiley, 736 pp
- Ehrlich R, Kennedy SK, Crabtree SJ, Cannon RL (1984) Petrographic image analysis, I. Analysis of reservoir pore complexes. *J Sediment Petrol* 54:1365–1376
- Ehrlich R, Full WE (1984a) Optimal definition of class intervals of histograms or frequency plots. In: Beddow JK (ed) Particle characterization in technology. CRC Press, Inc., Boca Raton, Florida, pp 135–148
- Ehrlich R, Full WE (1984b) Fourier shape analysis. In: Brown RM (ed) Pattern analysis in the marine environment. PAME Proceedings, pp 47–68
- Ehrlich R, Cobaleda G, Ferm JB (1997) Relationship between petrographic pore types and core measurements in sandstones of the Monserrate Formation, Upper Magdalena Valley, Colombia. *C.T.F Cienc. Tecnol. Futuro*, pp 5–17
- Ehrlich R, Horkowitz KO (1984) A strong transfer function links thin section data to reservoir physics. In: Proceedings of the 59th annual technical conference and exhibition, Society of Petroleum Engineers of AIME/SPE-13263-MS, 7 pp
- Ehrlich R, Horkowitz KO, Horkowitz JP, Crabtree SJ Jr (1991a) Petrography and reservoir physics I: objective classification of reservoir porosity. *Am Assoc Pet Geol Bull* 75:1547–1562
- Ehrlich R, Etris EL, Brumfield D, Yuan LP, Crabtree SJ Jr (1991b) Petrography and reservoir physics III: physical models for permeability and formation factor. *Am Assoc Pet Geol Bull* 75:1579–1592
- Ehrlich R, Wenning RJ, Johnson GW, Su SH, Paustenbach DJ (1994) A mixing model for polychlorinated dibenzo-*p*-dioxins and dibenzofurans in surface sediments from Newark Bay, New Jersey using polytopic vector analysis. *Arch Environ Contam Toxicol* 27:486–500

- Ehrlich R, Etris EL, Crabtree SJ (1996) Physical relevance of pore types derived from thin section by petrographic image analysis. In: SCA international symposium on core analysis. Dallas, Texas, vol 1, no 90001, 28 pp
- Ehrlich R, Full WE (1988) Sorting out geology-unmixing mixtures. In: Size W (ed) IAMG special publication. Oxford University Press, New York, pp 33–46
- El-Awawdeh RT, Full WE (1996) Quantification of coastline and key morphology changes over time in Northern Florida Bay, using high resolution shape analysis and aerial photographs. In: Proceedings of the second international airborne remote sensing conference, Environmental Research Institute of Michigan (ERIM), Ann Arbor, MI, Part I, pp 419–428
- Etris EL, Brumfield DS, Ehrlich R, Crabtree SJ Jr (1988) Relations between pores, throats and permeability: a petrographic/physical analysis of some carbonate grainstones and packstones. *Carbonates Evaporites* 3:17–32
- Evangelista S, Full WE, Tortora P (1986) Analisi Della Forma Particelle Sedimentarie: Una Applicazione ai Clasti Sabbiosi Di Ambrente Costiero. *Mem Soc Geol It* 35:823–826
- Evangelista S, Full WE, Tortora P (1994) Fourier grain shape analysis as tool to quantify the contribution of the fluvial input to the coastal sedimentary budget: an example from the Port Stephen's area, New South Wales, Australia. *Bollettino della Societa Geologica Italiana* 113:729–747
- Evangelista S, Full WE, Tortora P (1996) Contribution and dispersion of fluvial, beach, and shelf sands in the Bassa Maremma coastal system (Central Italy, Tyrrhenian Margin). *Bollettino della Societa Geologica Italiana* 116:195–217
- Evans JC, Ehrlich R, Krantz D, Full WE (1992) A comparison between polytopic vector analysis and empirical orthogonal function analysis for analyzing quasi-geostrophic potential vorticity. *J Geophys Res* Green 97:2365–2378
- Ferm JB, Ehrlich R, Crawford GA (1993) Petrographic image analysis and petrophysics: analysis of crystalline carbonates from the Permian basin, west Texas. *Carbonates Evaporites* 8:90–108
- Fillon RH, Full WE (1984) Grain size variations of North Atlantic non-carbonate sediments and sources of terrigenous components. *Mar Geol* 59:13–15
- Full WE, Ehrlich R, Kennedy SK (1984) Optimal configuration and information content of sets of frequency distributions. *J Sediment Petrol* 54:117–126
- Full WE, Ehrlich R, Klován JE (1981) EXTENDED QMODEL-Objective definition of external end members in the analysis of mixtures. *Math Geol* 13:331–344
- Full WE (1981) Analysis of quartz detritus of complex provenance via analysis of shape. PhD dissertation, Department of Geological Sciences, University of South Carolina, Columbia, SC, 206 pp
- Full WE (1987) The use of image analysis in petrography. In: Proceedings of electronic imaging '87, Boston, MA, pp 597–602
- Full WE, James S (2015) Rapid exploration in a mature area in northwest Kansas: improving the definition of key reservoir characteristics using big data, search and discovery article # 41685, 30 pp
- Full WE, Ehrlich R, Bezdek JC (1982) FUZZY QMODEL—a new approach for linear unmixing. *Math Geol* 14:259–270
- Gary AC, Johnson GW, Ekart DD, Platon E, Wakefield MI (2005) A method for two-well correlation using multivariate biostratigraphical data. In: Powell AJ, Riding JB (eds) Recent developments in applied biostratigraphy, pp 205–218
- Golub GH, Reinsch C (1970) Singular value decomposition and least squares solutions. *Numer Math* 14:403–420
- Healy-Williams N, Ehrlich R, Full W (1997) Closed-form Fourier analysis: a procedure for extracting ecological information from foraminiferal test morphology. In: Fourier descriptors and the applications in biology. Cambridge University Press, Cambridge, pp 129–156
- Healy-Williams N (1983) Fourier shape analysis of *Globorotalia truncatulinoides* from late quaternary sediments in the southern Indian Ocean. *Mar Micropaleontol* 8:1–15
- Healy-Williams N (1984) Quantitative image analysis: application to planktonic foraminiferal paleoecology and evolution. *Geobios Mémoire Spécial* 8:425–432

- Henk M, Richter-Gebert J, Ziegler GM (1997) Basic properties of convex polytopes. In: Goodman JE, O'Rourke J (eds) Handbook of discrete and computational geometry. CRC Press, Boca Raton, Chapter 19, pp 243–270
- Horkowitz J, Stakes D, Ehrlich R (1989) Unmixing mid-ocean ridge basalts with EXTENDED QMODEL. *Tectonophysics* 165:42754
- Hudson CB, Ehrlich R (1980) Determination of relative provenance contributions in samples of quartz sand using Q-mode factor analysis of Fourier grain shape data. *J Sediment Petrol* 50:1101–1110
- Huntley SL, Carlson-Lynch H, Johnson GW, Paustenbach DJ, Finley BL (1998) Identification of historical PCDD/F sources in Newark Bay Estuary subsurface sediments using polytopic vector analysis and radioactive dating techniques. *Chemosphere* 36:1167–1185
- Imbrie J (1963) Factor and vector analysis programs for analyzing geologic data, Office of Naval Research, Technical Report 6, Geography Branch, 83 pp
- James RA (1995) Application of petrographic image analysis to the characterization of fluid-flow pathways in a highly-cemented reservoir: Kane Field, Pennsylvania, USA. *J Pet Sci Eng* 13:141–154
- Jarman WM, Johnson GW, Bacon CE, Davis JA, Risebrough RW, Ramer R (1997) Levels and patterns of polychlorinated biphenyls in water collected from the San Francisco Bay and Estuary, 1993–1995. *Fresenius' J Anal Chem* 359:254–260
- JMP Pro (1989–2017) JMP Pro software package. SAS Institute Inc., Cary, NC
- Johnson GW, Ehrlich R, Full W (2001) Principal component analysis and receptor models in environmental forensics. In: Murphy BL, Morrison RD (eds) Introduction to environmental forensics. Academic Press, pp 461–515
- Johnson GW (1997a) Application of polytopic vector analysis to environmental geochemical investigation. PhD dissertation, University of South Carolina, 134 pp
- Johnson GW (1997b) Application of polytopic vector analysis to environmental geochemistry investigations. PhD dissertation, Department of Geological Sciences, University of South Carolina, Columbia, SC, 244 pp
- Johnson GW, Quensen JF III (2000) Implications of PCB dechlorination on linear mixing models. *Organohalogen Compd* 45:280–283
- Johnson GW, Jarman WM, Bacon CE, Davis JA, Ehrlich R, Risebrough R (2000) Resolving polychlorinated biphenyl source fingerprints in suspended particulate matter of San Francisco Bay. *Environ Sci Technol* 34:552–559
- Kaiser HF (1958) The varimax criterion for analytic rotation in factor analysis. *Psychometrika* 23:187–200
- Kensington E, Full WE (1994) Fourier analysis of scallop shells (*Placopecten magellanicus*) in determining population structure. *Can J Fish Aquat Sci* 51:348–356
- Klovan IE, Meisch AT (1976) EXTENDED CABFAC and QMODEL computer programs for Q-mode factor analysis of compositional data. *Comput Geosci* 1:161–178
- Klovan JE (1966) The use of factor analysis in determining depositional environments from grain-size distributions. *J Sediment Res* 36:115–125
- Klovan JE (1981) A generalization of extended 'Q'-mode factor analysis to data matrices with variable row sums. *J Int Assoc Math Geol* 13:217–224
- Klovan JE, Imbrie J (1971) An algorithm and Fortran-IV program for large-scale Q-mode factor analysis and calculation of factor scores. *Math Geol* 3:61–77
- Leather J, Durell D, Johnson G, Mills M (2012) Integrated Forensics approach to fingerprint PCB sources in sediments using rapid sediment characterization (RSC) and advanced chemical fingerprinting (ACF). Technical Document 3262, Final Report for the Environmental Security Technology Certification Program (ESTCP), 226 pp
- Leinen M, Piasias N (1984) An objective technique for determining end-member compositions and for partitioning sediments according to their sources. *Geochemica et Cosmochimica Acta* 48:47–62

- Lisowiec K, Slaby E, Förster HJ (2015) Polytopic vector analysis (PVA) modelling of whole-rock and apatite chemistry from the Karkonosze composite pluton (Poland, Czech Republic). *Lithos* 230:105–120
- Magar VS, Johnson GW, Brenner R, Durell G, Quensen JF III, Foote E, Ickes JA, Peven-McCarthy C (2005) Long-term recovery of PCB-contaminated sediments at the Lake Hartwell Superfund Site: PCB Dechlorination. I. End-member characterization. *Environ Sci Technol* 39:3538–3547
- Mason EW, Ehrlich R (1995) Automated, quantitative assessment of basin history from a multiwell analysis. *Am Assoc Pet Geol Bull* 79:711–724
- Mazzullo J, Ehrlich R (1980) A vertical pattern of variation in the St. Peter Sandstone Fourier grain shape analysis. *J Sediment Res* 50:63–70
- Mazzullo J, Ehrlich R (1983) Grain-shape variation in the St. Peter sandstone: a record of eolian and fluvial sedimentation of an early Paleozoic cratonic sheet sand. *J Sediment Res* 53:105–119
- Mazzullo J, Ehrlich R, Hemming MA (1984) Provenance and areal distribution of late Pleistocene and Holocene quartz sand on the southern New England continental shelf. *J Sediment Res* 54:1335–1348
- Mazzullo J, Ehrlich R, Pilkey OH (1982) Local and distal origin of sands in the Hatteras Abyssal Plain. *Mar Geol* 48:75–88
- McCreesh CA, Ehrlich R, Crabtree SJ (1991) Petrography and reservoir physics II: relating thin section porosity to capillary pressure, the association between pore types and throat size. *Am Assoc Pet Geol Bull* 75:1563–1578
- McKenna TE, Lerche I, Williams DF, Full WE (1988) Quantitative approaches to stratigraphic correlations and chemical stratigraphy. *Transactions, GCAGS/GCS-SEPM*, 38:128–135
- Megson D, Brown TA, Johnson GW, O'Sullivan G, Bicknell AWJ, Votier SC, Lohan MC, Comber S, Kalin R, Worsfold PJ (2014) Identifying the provenance of Leach's storm petrels in the North Atlantic using polychlorinated biphenyl signatures derived from comprehensive two-dimensional gas chromatography with time-of-flight mass spectrometry. *Chemosphere* 114:195–202
- Miesch AT (1976a) Q-mode factor analysis of geochemical and petrologic data matrices with constant row-sums. U.S. Geological Survey Professional Paper 574-G, 47 pp
- Miesch AT (1976b) Q-mode factor analysis of compositional data. *Comput Geosci* 1:147–159
- Mudge SM (2016) Statistical analysis of oil spill chemical composition data, 2nd edn. In: Stout S, Wang Z (eds) *Standard handbook oil spill environmental forensics*, pp 849–867
- Murillo-Jiménez JM, Full W, Nava-Sánchez E, Vera-Camacho V, León-Manilla A (2007) Sediment sources of beach sand from the southern coast of the Baja California Peninsula, México—Fourier grain shape analysis. *GSA Special Paper* 420:297–318
- Nash GD, Johnson GW (2002) Soil mineralogy anomaly detection in Dixie Valley, Nevada using hyperspectral data. In: *Proceedings of the twenty-seventh workshop on geothermal reservoir engineering*, Stanford University, Stanford, California, 28–30 Jan 2002 SGP-TR-171, pp 249–256
- Nash GD, Johnson GW, Johnson S (2004) Hyperspectral detection of geothermal system related soil mineralogy anomalies in Dixie Valley, Nevada: a tool for exploration. *Geothermics* 33:695–711
- Nelson RH Jr, Roberts JH, Siebert DJ, Massell WF, LeRoy SD, Denham LR, Ehrlich R, Coons RL (2013) Method for locating sub-surface natural resources. US Patent US 8344721 B2
- Porter GA, Ehrlich R, Osborne RH, Combellich RA (1979) Sources and nonsources of beach sand along southern Monterey Bay, California: Fourier shape analysis. *J Sediment Res Petrol* 49:727–732
- Reister DD, Shipp RC, Ehrlich R (1982) Patterns of quartz sand shape variation, Long Island littoral and shelf. *J Sediment Res* 52:1307–1314
- Ross CM, Ehrlich R (1991) Objective measurement and classification of microfabrics and their relationship to physical properties. In: Bennett RH et al (eds) *Microstructure of fine-grained sediments: from mud to shale*. Springer, pp 353–358

- Ross CM, Ehrlich R, Gellici JA, Richardson BO (1986) Relationship between wireline log response and quantitative pore geometry data in Shannon sandstone, Hartzog Draw Field, Wyoming. *Am Assoc Pet Geol Bull* 70:614–642
- Scheffe G, Full WE (1986) Petrographic image analysis of a sandstone reservoir in Oklahoma. In: *Proceedings of the Geotech symposium, Denver*, pp 196–202
- Schwartz JT (2011) *Introduction to matrices and vectors*. Dover Publications: Dover ed. Edition, 192 pp
- Shafer JL, Ehrlich R (1986) Method and apparatus for determining conformity of a predetermined shape related characteristics of an object or stream of objects by shape analysis. US Patent 4,624,367
- Smith MM, Ehrlich R, Ramirez R (1985) Quartz provenance changes through time: examples from two South Carolina barrier islands. *J Sediment Res* 55:483–494
- Smith SM (1997) National geochemical database reformatted data from the national uranium resource evaluation (NURE) hydrogeochemical and stream sediment reconnaissance (HSSR) program (No. 97-492). US Geological Survey, US Dept. of the Interior
- Solohub JT, Klován JE (1970) Evaluation of grain-size parameters in lacustrine environments. *J Sediment Petrol* 40:81–101
- Sophie B, Yannick A, Dominique D, Bertrand F, Fern JB (1999) Characterization by image analysis of the micro-geometry of the porous network of an illitized sandstone reservoir, Alwyn, North Sea. *Bulletin de la Societe Geologique de France* 170:367–377
- Stattegger K, Morton AC (1992) Enhancing tectonic and provenance information from detrital zircon studies: assessing terrane-scale sampling and grain-scale characterization. *J Geol Soc* 168:309–318
- Szymanski DW, Patino LC, Vogel TA, Alvarado GE (2013) Evaluating complex magma mixing via polytopic vector analysis (PVA) in the Papagayo Tuff, Northern Costa Rica: processes that form continental crust. *Geosciences* 3:585–615
- Tefend KS, Vogel TA, Flood TP, Ehrlich R (2007) Identifying relationships among silicic magma batches by polytopic vector analysis: a study of the Topopah Spring and Pah Canyon ash-flow sheets of the southwest Nevada volcanic field. *J Volcanol Geoth Res* 167:198–211
- Tortora P, Evangelista S, Full WE (1986) Fourier grain shape analysis and its application to the southeastern coastline of Australia. In: *National Geological Congress of Central Italy Proceedings, Rome, Italy*, pp 1–6
- Towey TP, Barabás N, Demond A, Franzblau A, Garabrant DH, Gillespie BW, Lepkowski J, Adriaens P (2012) Polytopic vector analysis of soil, dust, and serum samples to evaluate exposure sources of PCDD/FS. *Environ Toxicol Chem* 31:2191–2200
- Van de Wetering N, Mayer B, Sanei H (2015) Chemostratigraphic associations between trace elements and organic parameters within the Duvernay formation, Western Canadian Sedimentary Basin. In: *Unconventional Resources Technology Conference C: 2153955*, 13 pp
- Vogel TA, Hidalgo PJ, Patino L, Tefend KS, Ehrlich R (2008) Evaluation of magma mixing and fractional crystallization using whole-rock chemical analyses: polytopic vector analyses. *Geochem Geophys Geosyst*, vol 9. <https://doi.org/10.1029/2007gc001790>
- Weltje GJ (1997) End-member modeling of compositional data: numerical-statistical algorithms for solving the explicit mixing problem. *Math Geol* 29:503–549
- Wenning RJ, Erickson GA (1994) Interpretation and analysis of complex environmental data using chemometric methods. *TrAC Trends Anal Chem* 13:446–457
- Williams DF, Lerche I, Full WE (1988a) *Isotope chronostratigraphy: theory and methods*. Academic Press, 345 pp
- Williams DF, Ehrlich R, Spero HJ, Healy-Williams N, Gary AC (1988b) Shape and isotopic differences between conspecific foraminiferal morphotypes and resolution of paleoceanographic events. *Palaeogeogr Palaeoclimatol Palaeoecol* 64:153–162
- Yannick A, Bernard D, Ehrlich R (1996) Towards realistic flow modeling. Creation and evaluation of two-dimensional simulated porous media: an image analysis approach. *Surv Geophys* 17:265–287

- Zellers SD, Gary AC (2007) Unmixing foraminiferal assemblages: polytopic vector analysis applied to Yakataga formation sequences in the offshore Gulf of Alaska. *Palaios* 22:1443–1467
- Zhao GT, Wei Z, Full WE, Chen Q, Lin YS (2004) Fourier shape analysis and its application in geology. *Periodical of Ocean, University of China*, vol 34, pp 429–436

Open Access This chapter is licensed under the terms of the Creative Commons Attribution 4.0 International License (<http://creativecommons.org/licenses/by/4.0/>), which permits use, sharing, adaptation, distribution and reproduction in any medium or format, as long as you give appropriate credit to the original author(s) and the source, provide a link to the Creative Commons license and indicate if changes were made.

The images or other third party material in this chapter are included in the chapter's Creative Commons license, unless indicated otherwise in a credit line to the material. If material is not included in the chapter's Creative Commons license and your intended use is not permitted by statutory regulation or exceeds the permitted use, you will need to obtain permission directly from the copyright holder.

