AGU PUBLICATIONS

Geochemistry, Geophysics, Geosystems

RESEARCH ARTICLE

10.1002/2016GC006663

Key Points:
- To present statistical methods that effectively capture the structures including trends and groups inherent in geochemical data
- This approach is useful for any type of geochemical data, and we provide an Excel program file with which the readers can test the method
- We show how the isotopic compositions of basalts (MORB, OIB, arc, and continental basalts) are looked into with the methods

Supporting Information:
- Supporting Information S1
- Data Set S1 with Excel program KCA

Correspondence to:
H. Iwamori,
hikaru@jamstec.go.jp

Citation:
Iwamori, H., K. Yoshida, H. Nakamura, T. Kuwatani, M. Hamada, S. Haraguchi, and K. Ueki (2017), Classification of geochemical data based on multivariate statistical analyses: Complementary roles of cluster, principal component, and independent component analyses, *Geochem. Geophys. Geosyst.*, *18*, 994–1012, doi:10.1002/2016GC006663.

Received 29 SEP 2016
Accepted 22 JAN 2017
Accepted article online 28 JAN 2017
Published online 17 MAR 2017

Classification of geochemical data based on multivariate statistical analyses: Complementary roles of cluster, principal component, and independent component analyses

Hikaru Iwamori[1,2], Kenta Yoshida[1], Hitomi Nakamura[1,2,3], Tatsu Kuwatani[1], Morihisa Hamada[1], Satoru Haraguchi[1], and Kenta Ueki[4]

[1]Department of Solid Earth Geochemistry, Japan Agency for Marine-Earth Science and Technology, Natsushima, Yokosuka, Japan, [2]Department of Earth and Planetary Sciences, Tokyo Institute of Technology, Meguro, Tokyo, Japan, [3]Ocean Resources Research Center for Next Generation, Chiba Institute of Technology, Tsudanuma, Narashino, Chiba, Japan, [4]Earthquake Research Institute, The University of Tokyo, Bunkyo, Tokyo, Japan

Abstract Identifying the data structure including trends and groups/clusters in geochemical problems is essential to discuss the origin of sources and processes from the observed variability of data. An increasing number and high dimensionality of recent geochemical data require efficient and accurate multivariate statistical analysis methods. In this paper, we show the relationship and complementary roles of k-means cluster analysis (KCA), principal component analysis (PCA), and independent component analysis (ICA) to capture the true data structure. When the data are preprocessed by primary standardization (i.e., with the zero mean and normalized by the standard deviation), KCA and PCA provide essentially the same results, although the former returns the solution in a discretized space. When the data are preprocessed by whitening (i.e., normalized by eigenvalues along the principal components), KCA and ICA may identify a set of independent trends and groups, irrespective of the amplitude (power) of variance. As an example, basalt isotopic compositions have been analyzed with KCA on the whitened data, demonstrating clear rock type/tectonic occurrence/mantle end-member discrimination. Therefore, the combination of these methods, particularly KCA on whitened data, is useful to capture and discuss the data structure of various geochemical systems, for which an Excel program is provided.

Plain Language Summary This paper presents a new statistical method that effectively captures the structures of various types of multivariate data (not only geochemical data but also any type of data). The method is based on combinations of k-means cluster analysis, principal component analysis, and independent component analysis for preprocessed data. The corresponding Excel program file is provided.

1. Introduction

Continuous material differentiation and homogenization associated with various geological processes have occurred through the Earth's history. As a result, geochemical data derived from geological samples often show compositional trends and groups. In turn, by identifying such trends and groups in the samples and data, we may infer the sources and processes that produced the compositional variability. A classic example is the crystallization trends in igneous rock suites [*Bowen*, 1928; *Fenner*, 1929; *Miyashiro*, 1975], which produce a wide compositional range of the Earth's crust. Another example concerns the radiogenic isotope systems, e.g., Rb-Sr, Sm-Nd, and U-Th-Pb-He, based on which a part of the ocean island basalts (OIB) are grouped into several "mantle geochemical components" such as HIMU, FOZO/C, EM1, and EM2 [*Zindler and Hart*, 1986; *Hart et al*., 1992; *Hofmann*, 2003; *Stracke*, 2012]. Mid-ocean ridge basalts (MORB) from the three major oceans exhibit overlapping but specific characteristics and can be grouped into Pacific Ocean-type and Indian Ocean-type [*Hickey-Vargas*, 1998; *Hofmann*, 2003]. These groups and signatures are interpreted to represent various recycled components and geochemical domains in the mantle.

In most cases, the geochemical trends and groups have been graphically identified on the basis of a series of two-dimensional (2-D) diagrams with observed variables, such as elemental concentration, elemental

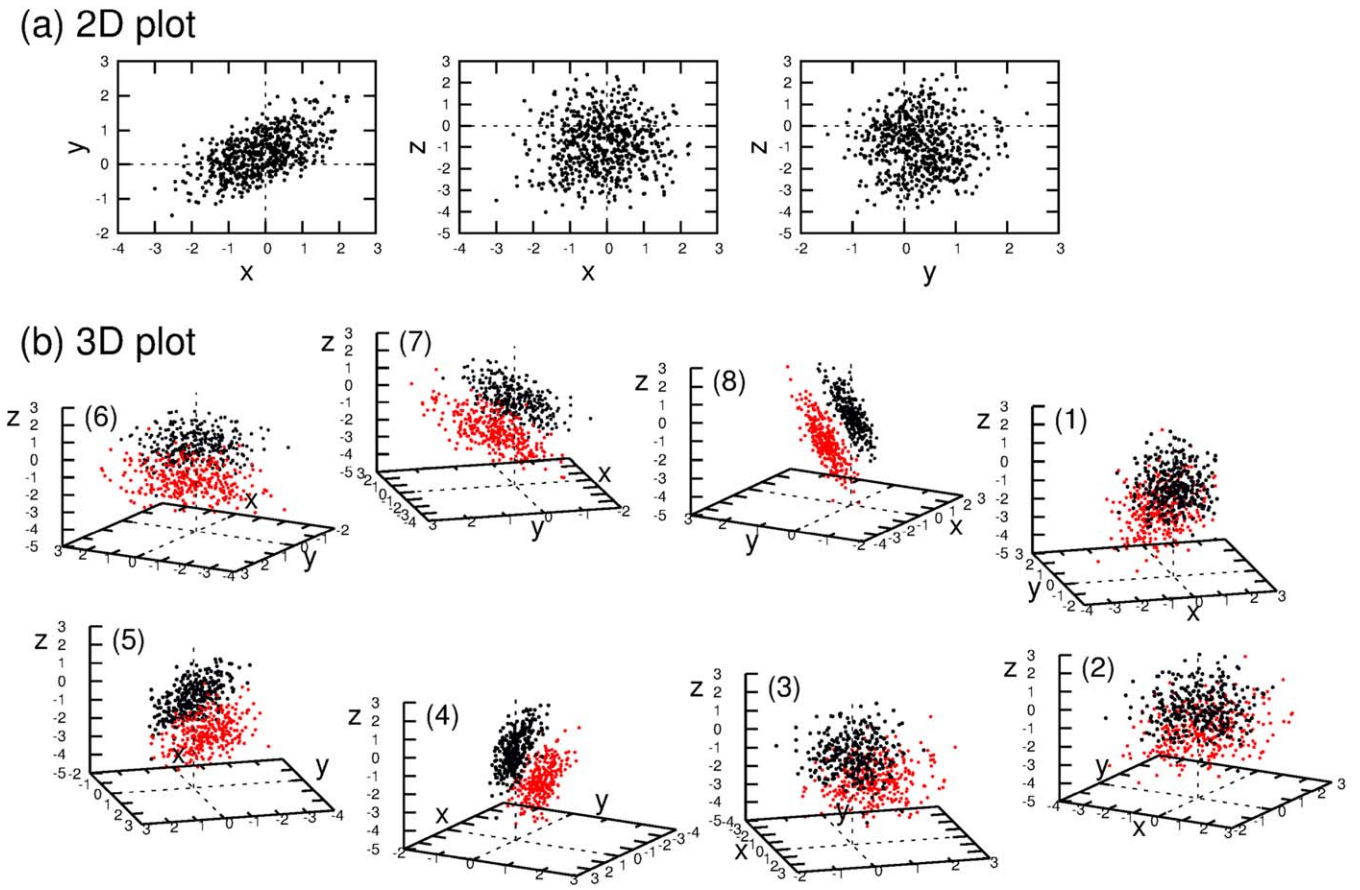

Figure 1. Plot of synthetic data with three variables $[x,\ y,\ z]$ from different view angles. (a) 2-D projections on $x-y$, $x-z$, and $y-z$ planes. (b) 3-D views. The data consist of two separate clusters. In Figure 1a, all of the data from two clusters are colored in black. In Figure 1b, the data from two clusters are color-coded in black and red.

ratio, isotopic ratio, and their combinations. The characteristics of data variability are captured in terms of coherence and structure among the variables such as overall correlations as well as the degrees of joint normality, lognormality, and bimodality to identify trends and groups. However, such a graphical approach is vulnerable for multivariate data sets with more than three dimensions. Figure 1 shows such an example with a synthetic three-dimensional (3-D) data set with the observed variables $[x,\ y,\ z]$. Figure 1a shows the data distribution on $x-y$, $x-z$, and $y-z$ planes. Only a broad positive correlation between x and y is seen, whereas almost no correlation is recognized on the $x-z$ and $y-z$ planes, although there are two distinct groups (Figure 1b). This type of coherent distribution commonly occurs in natural samples such as those from a melt-residue system, because a group of elements are controlled by the partition coefficients and behaves coherently. Accordingly, fewer opportunities are available for capturing the true structure in the observed variable space.

Detailed inspection of the 3-D views, as given in Figure 1b, can identify the data structure with three variables. However, this method is not applicable to systems with four or more variables. The two databases of GEOROC (http://georoc.mpch-mainz.gwdg.de/georoc/) and PetDB (http://www.petdb.org) [*Lehnert et al.*, 2000] contain ~382,000 sets of data in total. Of these, more than 25,000 sets of data contain 25 trace elements (commonly used for a spidergram), and more than 25,800 sets include five isotopic ratios of Sr, Nd, and Pb. *Jenner and O'Neil* [2012] provided analysis of 60 elements in 616 ocean floor basaltic glasses. The structure including trends and groups of these data cannot be identified by graphical methods. Even 2-D data may be misinterpreted by graphical methods, as will be demonstrated.

In order to correctly and fully extract information from multivariate geochemical data sets, statistical approaches are required for classification and clustering, and are commonly used in environmental sciences, chemometrics, and geochemical exploration [*Varmuza and Filzmoser*, 2009; *Reimann et al.*, 2011]. In

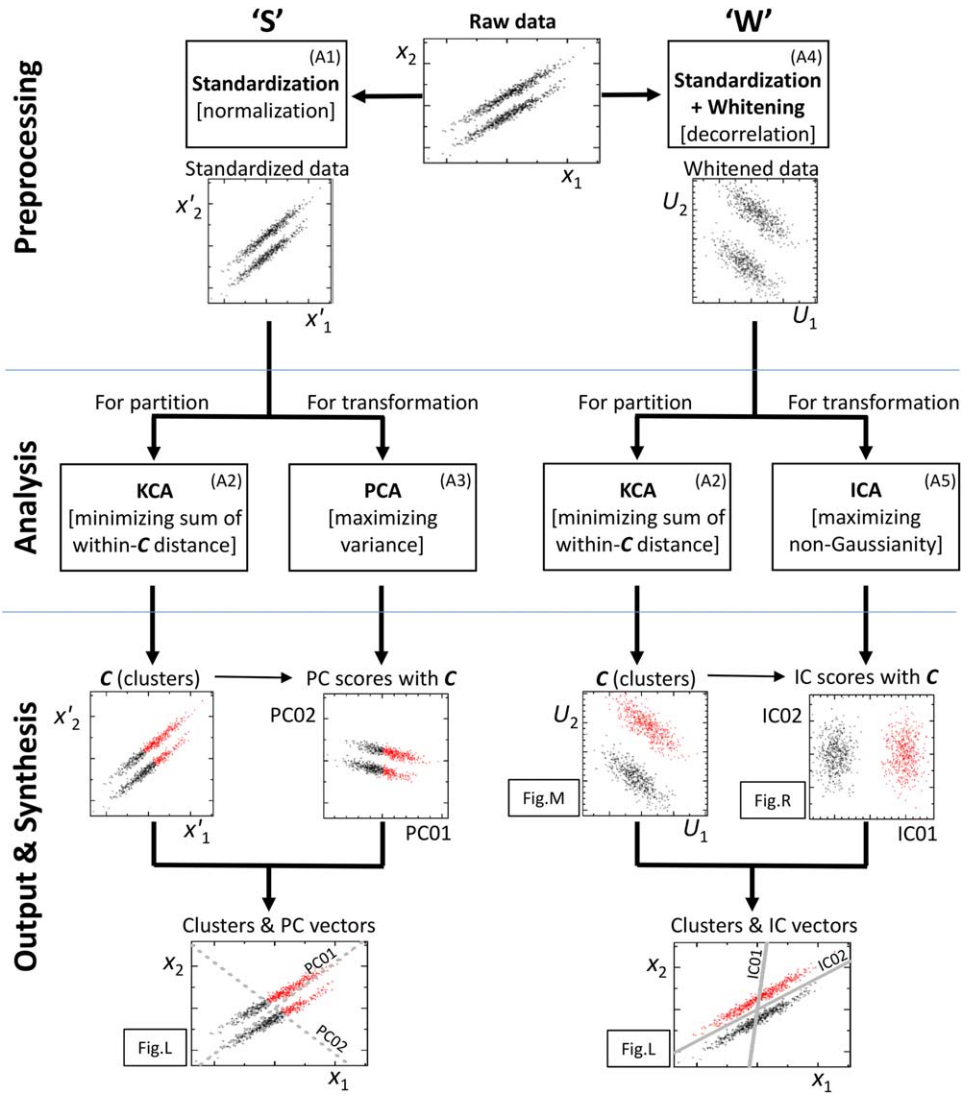

Figure 2. Schematic diagram showing outline and relationship of the methods used in this study. See section 2 for detail.

geochemistry, petrology-mineralogy, and geology, although such approaches seem slightly less common, various multivariate statistical methods have been applied to unravel the data structures including trends, groups, and related end-members. These include linear regression, cluster analysis, discriminant analysis, principal component analysis, factor analysis, and independent component analysis: e.g., application to various petrological problems [*Le Maitre*, 1982]; identification of mantle geochemical structures [*Zindler et al*., 1982; *Allègre et al*., 1987; *Hart et al*., 1992; *White and Duncan*, 1996; *Iwamori and Albarède*, 2008; *Stracke*, 2012]; classification and source identification of sediment [*Pisias et al*., 2013; *Yasukawa et al*., 2016] or volcanic rocks [*Brandmeier and Wörner*, 2016]; and rock-tectonic setting association [*Agrawal et al*., 2004; *Snow*, 2006; *Vermeesch*, 2006; *Verma et al*., 2013]. Additionally, advanced methods of supervised machine learning have been applied recently to identify Tsunami deposits [*Kuwatani et al*., 2014] and tectonic discrimination of igneous rocks based on PetDB and GEOROC databases [*Petrelli and Perugini*, 2016].

Here we systematically apply three statistical analyses to geochemical data and identify the structures including trends and groups. The three analyses are nonhierarchical cluster analysis (k-means cluster analysis, hereafter referred to as KCA), principal component analysis (PCA), and independent component analysis (ICA). These three methods are of an unsupervised approach, and do not require prior information to unravel the hidden structures based upon which classification/grouping can be made based solely on the data. KCA is a simple and commonly used method to partition the data into an assigned number of clusters,

where the data are classified into several discretized groups. PCA and ICA described the data by a set of base vectors that maximize either the variance or the non-Gaussianity, respectively, in which case the data are described in a continuous solution space with individual scores. Graphical and intuitive explanations of KCA, PCA, and ICA are shown in Figure 2, particularly in the bottom two diagrams: PCA finds the main base vector shown as PC01 in the bottom left diagram of Figure 2, along which the overall data distribution is the most elongated. ICA finds the structure, shown in the bottom right diagram as IC01 separating the two trends and IC02 parallel to the trends, inherent in the data. KCA partitions the data along the PCs or ICs, as will be described later in detail.

The reasons for selecting the three methods are (i) KCA and PCA are probably the most fundamental yet powerful tools for multivariate analyses; (ii) ICA is not as common as PCA but is a unique tool for identifying hidden independent structures; and (iii) the three methods are newly found to be closely related and can be integrated to analyze the data effectively. In this study, we describe the relationship of these three methods to elucidate the entire data structure based mainly on synthetic data, and we show the potential usefulness for geochemistry. We apply this to a natural data set of isotopic compositions of basalts for which ICA has been performed. On the basis of the results, an effective combination (in terms of computational procedure and resource) of the methods is clarified, for which we provide an Excel program (supporting information and http://dsap.jamstec.go.jp/) to allow readers to test and apply the program to individual problems.

2. Methods

An outline of the methods is schematically illustrated in Figure 2, and the corresponding equations are described in Appendix. The methods consist of three steps. First, the data are preprocessed (standardized or standardized + whitened; "Preprocessing" in Figure 2), and then analyzed by KCA, PCA, and ICA ("Analysis" in Figure 2). Finally, the outputs are plotted and synthesized in several forms with different variables and components for interpretation and visualization ("Output & Synthesis").

The observed variables on a suite of samples generally have different ranges and units. First, a primary standardization of raw data $\mathbf{X}$ using the mean and standard deviation is applied (equation (A1) in Appendix), where $\mathbf{X} = (x_{ij})$ is the data matrix for the ith sample and the jth element/isotope, to avoid the effects of such differences on the statistical results. Several other standardization procedures, e.g., using the median or median absolute deviation instead of the mean, can be applied to reduce the effects of outliers. For geochemical data, analytical uncertainties in principle should be incorporated, in addition to normalization, by the standard deviation. If the data are under the constant-sum constraint, such as those represented by percent or parts per million, then preprocessing of the raw data by taking the logratio (additive logratio transformation or centered logratio transformation) is a useful method [*Aitchison*, 1986], instead of or in addition to the primary standardization. In this study, we use the primary standardization only with the mean and standard deviation for simplicity.

The standardized data $\mathbf{X}'$ are then analyzed by KCA and PCA as shown under column "S" in Figure 2. KCA is a widely used classification method for partitioning the multivariate data into a set of K clusters $\boldsymbol{C} = \{c_k\}_{k=1,\ldots,K}$, such that the total distance between the mean of a cluster (centroid) and the individual data points in the cluster (within-$\boldsymbol{C}$ distance) is minimized as expressed by equation (A2) in Appendix [*MacQueen*, 1967; *Jain*, 2010]. In KCA, the users assign K. In Figure 2, $K = 2$ is assumed and the data are partitioned into the two clusters colored by black and red. At the start of the KCA calculation, an arbitrary or estimated set of cluster centroids is assumed, which determine the centroid closest to the individual data points. Then the revised centroids are calculated for each cluster and are used to again determine the new clusters and centroids. This procedure repeats until the centroids become stationary. The standard iterative solution to KCA may be trapped in the local minima [*Bradley and Fayyad*, 1998]. Accordingly, a total of 100,000 trials was performed under randomly different initial conditions to find the global minimum in this study for the number of samples $n = 1000$ to $10,000$, although several algorithms can be used to improve the efficiency (e.g., k-means++) [*Arthur*, 2007].

PCA is also a commonly used method for specifying the uncorrelated base vectors that effectively account for the data by maximizing the variance along the principal components (PCs). The eigenvalue decomposition of the variance-covariance matrix for standardized data (i.e., the correlation coefficient matrix for the raw data) gives eigenvectors, such as PC vectors or "factor loadings" of individual PCs, and the corresponding eigenvalues. Then, the data are expressed in the PC space as "PC scores" (equation (A3) in Appendix). The first PC (PC01) corresponds to the overall elongation of data along which the sample variance is maximum. In Figure 2, the diagonal direction of two trends maximizes the variance and corresponds to PC01;

hence, the PCs and the two trends are oblique. The PC scores in Figure 2 are color-coded according to the clustering, as shown in the subplot titled "PC scores with $\boldsymbol{C}$." The obtained results, e.g., cluster indices and PC vectors, can be plotted in the original variable space, as shown in the bottom left subplot of Figure 2.

In the right-hand column "W" of Figure 2, "Whitening" (or "Sphering") is applied in addition to the standardization in order to decorrelate the variables. The whitened (or sphered) data $\mathbf{U}_r$ are obtained by dividing the PC scores by the square root of the eigenvalues as expressed by equation (A4) in Appendix. Then, any orthogonal pair of the base vectors is uncorrelated and the variances are equalized; therefore, an ellipsoidal data distribution is adjusted to a spherical distribution, termed "Sphering." In this case, instead of the variance, or the amplitude or power of the variations, the directional structures become highlighted and important, and $\mathbf{U}_r$ can be analyzed by KCA and ICA. The dimension of $\mathbf{U}_r$ can be reduced on the basis of the eigenvalues, and the PCs with sufficiently large eigenvalues can be selected. This brings several advantages such as computation saving, noise removal, calculation stabilization; however, important information for KCA and ICA may be lost. In this study, the data are not reduced, partly because $m=2$ or 3 only for the synthetic data in this study. If characteristic structures inherent in the data are expected to be disturbed significantly by noise, dimension reduction of this data is recommended.

Compared to KCA and PCA, ICA is a less common but powerful method used to extract independent base vectors and has been employed to solve various problems in information and brain sciences [*Hyvärinen et al*., 2001], including informatics of geochemical data [*Iwamori and Albarède*, 2008; *Iwamori et al*., 2010]. Two random variables are independent if their joint probability density function (PDF) equals the product of two PDFs of the individual variables [*Hamilton*, 1964]. The principle of ICA is straightforward: If the original independent source signals exhibit non-Gaussian distribution, such as homogeneous or skewed distribution, a signal that randomly mixes the independent sources shows a distribution closer to the Gaussian distribution. In turn, a linear combination of the observed mixture variables will be maximally non-Gaussian if it equals one of the independent components (ICs) [*Hyvärinen et al*., 2001]. The ICs or IC vectors can be found by rotating an orthogonal pair of base vectors to maximize the non-Gaussianity within the whitened space; ICA is always performed on the whitened data. In the right-hand column of Figure 2, the whitened data include two broad trends that are aligned parallel in the IC space (subplot titled "IC scores with $\boldsymbol{C}$"), indicating that IC01 and IC02 have no mutual information on the other. In this study, we used the FastICA algorithm [*Hyvärinen*, 1999] to search for the ICs by employing negentropy as a measure for non-Gaussianity to hold robustness against outliers [*Hyvärinen*, 1999; *Iwamori and Albarède*, 2008].

ICA has a few but key limitations [*Hyvärinen et al*., 2001; *Iwamori et al*., 2010]. First, the ICs must have non-Gaussian distributions to be detected. If the ICs have Gaussian distributions, ICA fails to identify the ICs because they have zero non-Gaussianity and cannot be discriminated from other nonindependent components. A second key point is that the user of ICA specifies the number of ICs to be detected, up to a maximum of the dimension of data. In this study, we first whitened the data by using PCA as described above. On the basis of the eigenvalues, as is commonly assumed in both PCA and ICA, components that account for a small proportion of the variance are judged to be unimportant signals including noise. This process is known as dimension reduction and determines the number of ICs, which in many cases stabilizes ICA by reducing noise [*Hyvärinen et al*., 2001]. Therefore, the number of IC is determined by a compromise between dimension reduction and retainment of the original information.

In section 3, these methods are first applied to a series of synthetic data for understanding the characteristics of KCA, PCA, and ICA, and are then applied to a natural system as an example to demonstrate how the actual data can be analyzed with these methods. The number of data points is an important factor for statistical analyses. In principle, more data points result in better results. In actual use, however, it depends on the data structure. For example, ICA for five-dimensional isotopic data of basalts requires several hundred or more samples to correctly capture the independent structures [*Iwamori and Albarède*, 2008]. In this study, the main purpose is to understand the behavior of and the relationship between PCA, ICA, and KCA. We used a sufficiently large number of data: approximately 1000 data points for the individual synthetic system and 6854 for the natural system.

3. Results

To systematically describe the results of KCA, PCA, and ICA for the synthetic data, a suite of three figures that correspond to "Fig. L," "Fig. M," and "Fig. R" in Figure 2 are shown in the left, middle, and right columns,

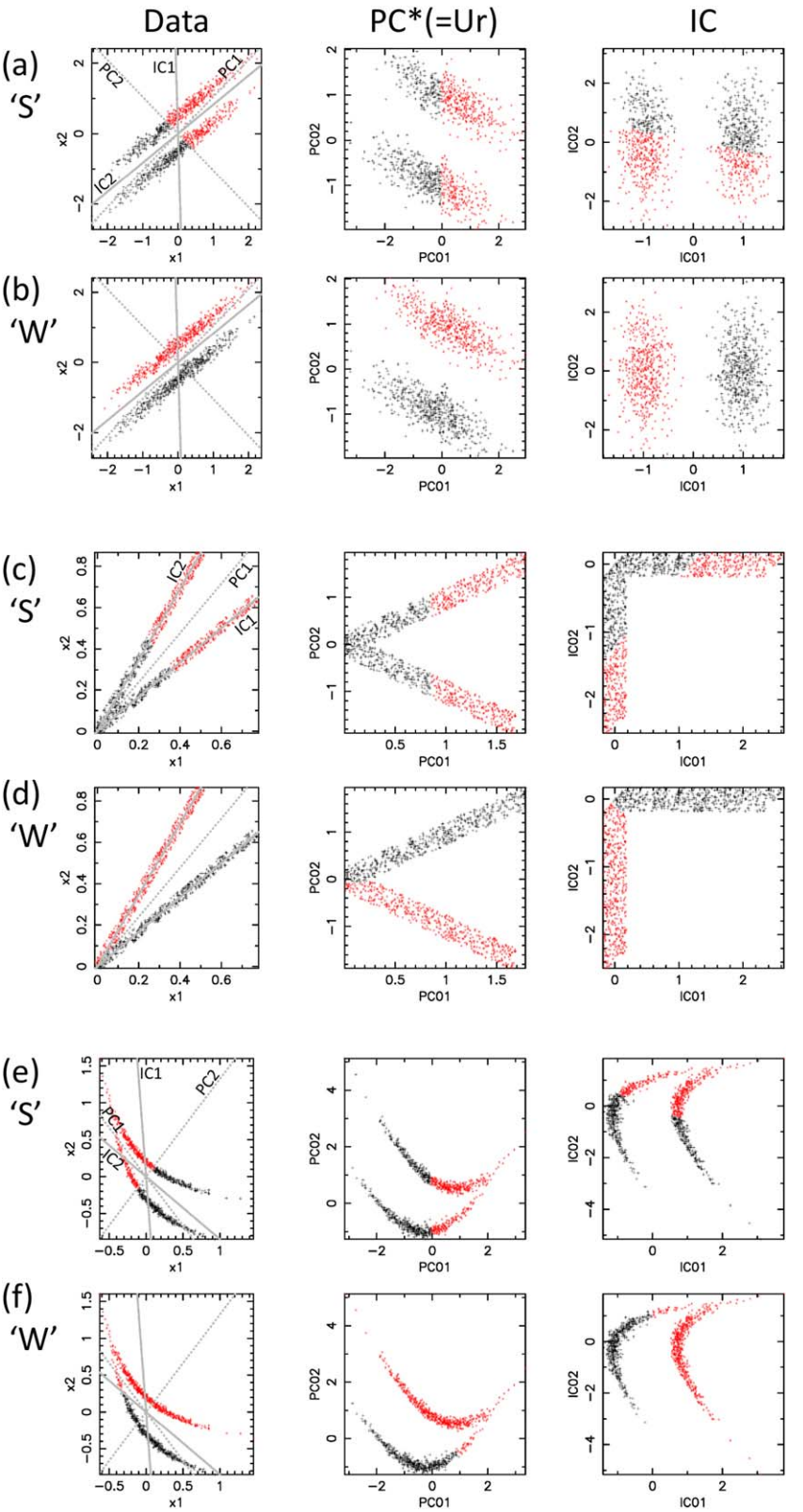
Data
PC*(=Ur)
IC
(a)
'S'
(b)
'W'
(c)
'S'
(d)
'W'
(e)
'S'
(f)
'W'
IC1
IC2
PC1
PC2
x1
x2
PC01
PC02
IC01
IC02

Figure 3.

respectively, in Figures 3–6. In the left column, both PCs and ICs are shown together with clusters in the original variable space. In the middle column, the whitened data $\mathbf{U}_r$ which are equivalent to eigenvalue-normalized PC scores, are shown as PC*. It is noted that PC* represents the results of PCA, having an essentially similar structure to the PC scores; therefore, they are referred to as PC in the following arguments. These values can be compared directly with the IC in the same whitened space. The right-hand column shows the IC scores with clusters.

The left-hand column in Figures 3a–3f shows three types of synthetic data with two variables x_1 and x_2, each consisting of two separate or converged trends. Such trends may occur in association with multiple binary mixing or distinct fractionation events from a single source. The two parallel trends in Figures 3a and 3b are identical, having the same two principal components (PC1 and PC2, shown as dashed lines) and the two independent components (IC1 and IC2, shown as solid lines) as in the left column. Although $K=2$ (two clusters) are assumed for both Figures 3a and 3b, the results differ depending on the preprocessing of the data: (a) is based on the primary standardized data ($\mathbf{X}'$, equation (A1) in Appendix, the left-hand column "S" in Figure 2); whereas (b) is based on the whitened data ($\mathbf{U}_r$, equation (A4), the right-hand column "W" in Figure 2). In Figure 3a, the overall elongated distribution along $x_1 \sim x_2$ is divided in half, and the boundary between the two clusters (black and red) coincides with PC2 at PC1=0 (left and middle columns of Figure 3a). PC1 is slightly but clearly oblique to the two parallel trends, because the maximum variance occurs along a diagonal direction oblique to the two trends, along the PC1 direction. In this case, PC1 and PC2 are not independent (middle column), and fail to separate the two trends; that is, the trends cannot be identified by either PC alone. On the contrary, the two ICs separate the trends (left and right columns): IC2 is parallel to the trends and IC1 measures the distance between the two trends, both being independent of each other.

In both the middle and right columns, the whitened data are plotted but with different rotation angles according to either the PCs or ICs, showing more isotropic distribution than the data in the left column. In Figure 3b, clustering has been performed on the whitened data, resulting in clear separation of the two trends, unlike that in Figure 3a. Similar behavior was observed for the other two cases of Figures 3c and 3d and Figures 3e and 3f, respectively. In Figure 3c with primary standardization, each of the two oblique trends diverging from the origin is partitioned into two parts along each trend, whereas the two trends are clearly separated in Figure 3d by clustering based on the whitened data. In both Figures 3c and 3d, PC1 corresponds to the average slope between the two trends, whereas the two ICs coincide with the individual trends. In the two cases of Figures 3a and 3b and Figures 3c and 3d, ICA discriminates the independent groups and trends, whereas PCA fails to identify these features and mixes originally independent information, as was noted by *Hyvärinen et al*. [2001]. Even for nonlinear curved trends in Figures 3e and 3f, which might resemble fractional crystallization paths, e.g., those on an SiO_2 versus MgO diagram [*Nakamura and Iwamori*, 2013], ICA helps to some extent to identify independent features, although the nonlinear trends are not supposed to be properly captured by the linear PCA and ICA of this study. The two parts with different overall slopes for two trends are broadly identified in the right-hand column of Figures 3e and 3f. The KCA results were also affected by the nonlinearity because KCA and PCA/ICA share common features, as will be discussed subsequently. However, clustering with the whitened data may broadly separate the two curved trends except for those in the crossing area (Figure 3f).

The presence of two PC or IC axes suggests that the data may be clustered into four quadrants in 2-D space. Figure 4 shows the results of KCA when the number of clusters (K) is given as 4. It is noted that the results

Figure 3. Analyses of three types of synthetic data, [(a), (b)], [(c), (d)], and [(e), (f)], with two variables. In all (a)–(f), the left column plots the 1000 raw data with PCs and ICs as in "Fig. L" of Figure 2; the middle column shows the whitened data $\mathbf{U}_r$ (= PC*, normalized PC scores, see the main text for detail) as in "Fig. M" of Figure 2; and the right column plots the IC scores as in "Fig. R" of Figure 2. In each diagram, the results of KCA with the cluster number $K=2$ are shown by color coding. In Figure 3a, a two-variable joint Gaussian distribution (with the zero mean and the standard deviations of unity and 0.1) is rotated by 40°. Another joint Gaussian distribution is created with the same procedure and is placed at +0.5 in terms of the vertical coordinate, producing the two parallel trends. Both Figures 3a and 3b use exactly the same raw data and therefore, the same PCs and ICs, but are different in the KCA results depending on the preprocessing either by primary standardization "S" in Figure 3a, or whitening "W" in Figure 3b, for KCA. In Figure 3c, two rectangular joint homogeneous distributions in the diagonal range of (0.0, −0.05) and (1.0, 0.05) are rotated by 40° and 60°, respectively, crossing at the origin. In this case, PCA and ICA have been performed around the origin instead of the sample mean. In Figure 3d, the same data as in Figure 3c are used, and the difference is the preprocessing of "S" in Figure 3c and "W" in Figure 3d for KCA. In Figure 3e, the same joint Gaussian distribution (x, y) as in Figure 3a is used to calculate $(x', y') = \left(1.5^{(x\cos a - y\sin a)} - b, 1.5^{-(y\cos a + x\sin a)} - c\right)$ where $(a, b, c) = (42°, -1.0, -0.8)$ and $(50°, -1.0, -1.3)$ to produce the two curves. In Figure 3f, the same data as in Figure 3e are used, and the difference is the preprocessing of "S" in Figure 3e and "W" in Figure 3f for KCA.

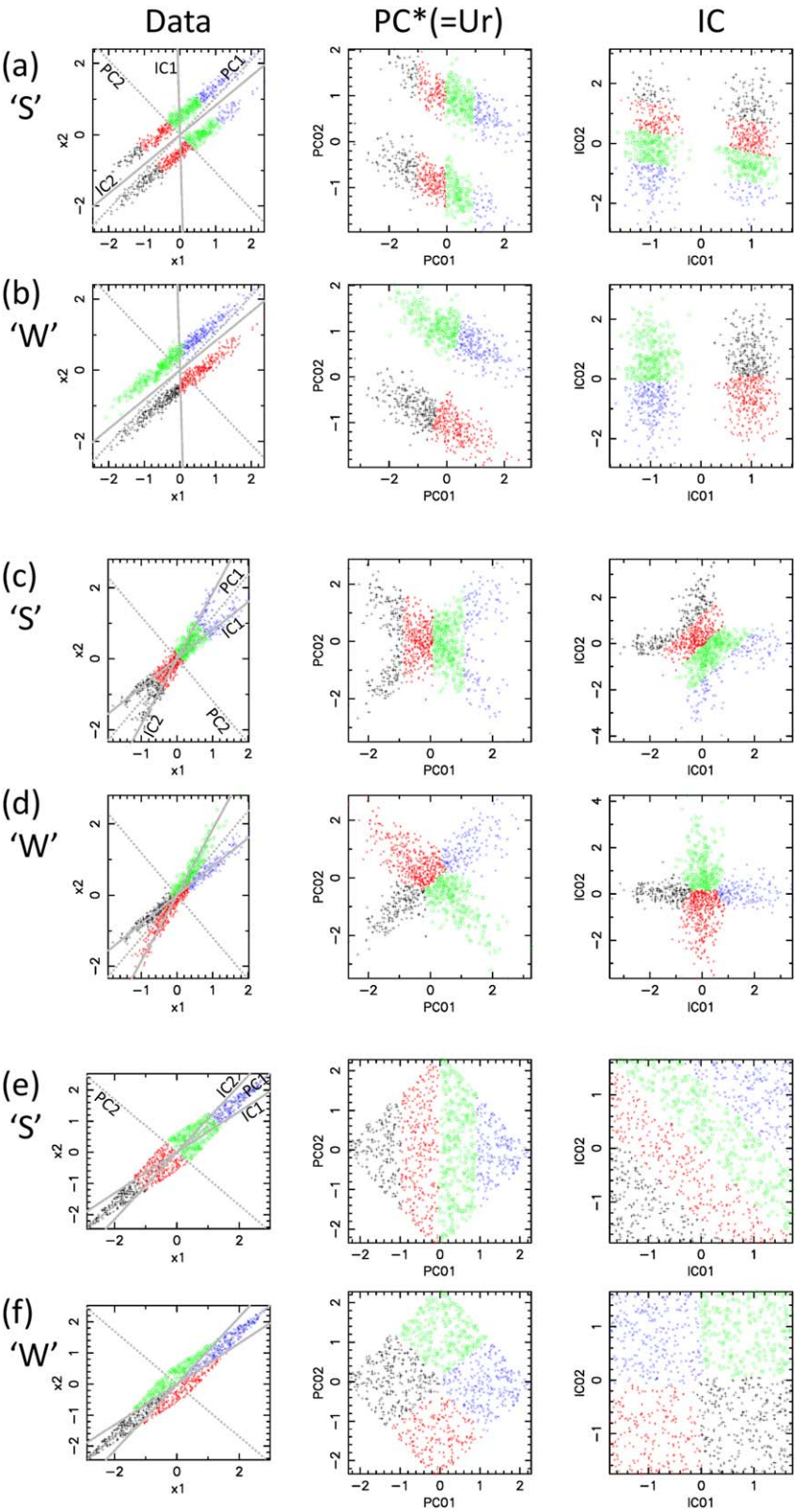
Data
PC*(=Ur)
IC
(a)
'S'
(b)
'W'
(c)
'S'
(d)
'W'
(e)
'S'
(f)
'W'
PC1
PC2
IC1
IC2
x1
x2
PC01
PC02
IC01
IC02

Figure 4.

of PCA and ICA are not affected by the number of clusters. Figures 4a and 4b uses the same data as those in Figures 3a and 3b, and the behavior described for Figures 3a and 3b is even clearer. In Figure 4a, the data on the individual trends are divided along the PC1 direction into four parts (left and middle columns of (a)). In Figure 4b, the data are sharply partitioned into four quadrants in the IC-coordinates (right columns), indicating the affinity between KCA with the whitened data and ICA. To confirm the affinity, two more types of data, each synthesized on the basis of two independent base vectors, have been analyzed, as shown in Figures 4c and 4d and Figures 4e and 4f. Each uses data that are different from each other and also different from Figures 3c and 3d and Figures 3e and 3f, respectively. The two crossing trends are divided into four parts along PC1 in Figure 4c with primary standardization. In Figure 4d with the whitened data, four branches stemming from the mean are identified and separated by KCA and ICA. Figures 4e and 4f show homogeneous distribution spanned by the two base vectors with directions parallel to the edges of the elongated parallelogram. This type of distribution is actually observed for the basalt isotopic composition in terms of the $^{206}Pb/^{204}Pb$ versus $^{208}Pb/^{204}Pb$ diagram, due to slightly oblique base vectors that might correspond to two different parent/daughter fractionation processes [*Iwamori and Nakamura*, 2015]. The results shown in Figure 4f might look awkward intuitively, but they are found to be reasonable when plotted in the IC space: The four quadrants in IC1-IC2 exactly coincide with the four clusters.

Further increase in the number of clusters is found to provide better resolution for the data structure, especially for the central part where the number density of data is maximum, as is common in many cases (Figures 5a–5d). For example, the central part and the surrounding four branches are separated by assigning $K=5$ for clustering the whitened data (Figure 5b), whereas five clusters are simply aligned along PC1 for primary standardization (Figure 5a). If three trends overlap, the crossing point (although not necessarily a point as in Figures 5c and 5d) may have even higher number density of the data. When $K=9$ for clustering the whitened data, the resultant structure involves six branches with the three clusters in the crossover area. Each cluster in the crossover area corresponds to a root part of the two branches (Figure 5d), whereas the clusters again are simply aligned mostly along PC1 in Figure 5c. It is noted that not only PCA but also ICA cannot correctly resolve the structure in this case; resolving three trends requires at least three variables. The ICs are not completely parallel in two of the three trends but show some obliquity depending on the symmetricity of the data distribution (Figures 5c and 5d). However, KCA works to discriminate the trends because it is based simply on the distance between the data points irrespective of the dimensions of variables.

All arguments made above for 2-D cases can be extended to higher dimensions. Figure 6 shows the 3-D case with three crossing trends in the three variable space. The left column shows the raw data, and ICA exactly identifies the three trends (right column), whereas PCA returns PCs that are oblique to the trends (middle column). The differences between Figure 6a for the standardized data and 6b for the whitened data are essentially similar to that in the 2-D cases: The clusters are aligned along the PCs in 6a, whereas the six branches are discriminated with a center part cluster in 6b. High-dimensional data for major element compositions of polymictic regoliths represented by mixtures of howardite, eucrite, and diogenite (HED) meteorites constitute multiple trends crossing each other [*Usui and Iwamori*, 2013] and are broadly similar to the variation in Figure 6. Although KCA has not been applied to the data set, ICA has resolved the mixing relationships [*Usui and Iwamori*, 2013]; therefore, KCA on the whitened data is expected to be useful for classifying the meteorites.

In the examples shown, Figures 3–6 are variably synthesized using transformation and a combination of Gaussian or homogeneous distribution, as described in the figure captions. The distribution type in each trend or cluster is not critical to the arguments above because the overall structures, such as multiple trends or clusters, are dominant. The synthesized data sets are given in the Excel program file "KCA" in supporting information and at http://dsap.jamstec.go.jp/, and can be re-examined with the included Excel program.

Figure 4. Analyses of three types of synthetic data, [(a), (b)], [(c), (d)], and [(e), (f)], with two variables. In all (a)–(f), the left column plots the 1000 raw data with PCs and ICs as in "Fig. L" of Figure 2; the middle column shows the whitened data **U,** (=PC*, normalized PC scores) as in "Fig. M" of Figure 2; and the right column plots the IC scores as in "Fig. R" of Figure 2. In each diagram, the results of KCA with the cluster number $K=4$ are shown by color coding. In Figure 4a, the same data as in Figure 3a are used. Accordingly, the PCs and ICs are the same. The difference between Figure 4a and Figure 4b is the preprocessing of "S" (primary standardization) in Figure 4a and "W" (whitening) in Figure 4b for KCA. In Figure 4c, two joint Gaussian distributions as in Figure 3a are rotated 40° and 60°, respectively, producing two crossing trends. In Figure 4d, the same data as in Figure 4c are used, and the difference is the preprocessing of "S" in Figure 4c and "W" in Figure 4d for KCA. In Figure 4e, a joint homogeneous distribution (x, y) in the diagonal range of $(-1.0, -0.1)$ and $(1.0,\ 0.1)$ is transformed as $(x', y')=\left(\frac{x+y}{2}, \frac{x-y}{4}\right)$ and is then rotated by 40°. In Figure 4f, the same data as in Figure 4e are used, and the difference between Figure 4e and Figure 4f is the preprocessing of "S" in Figure 4e and "W" in Figure 4f for KCA.

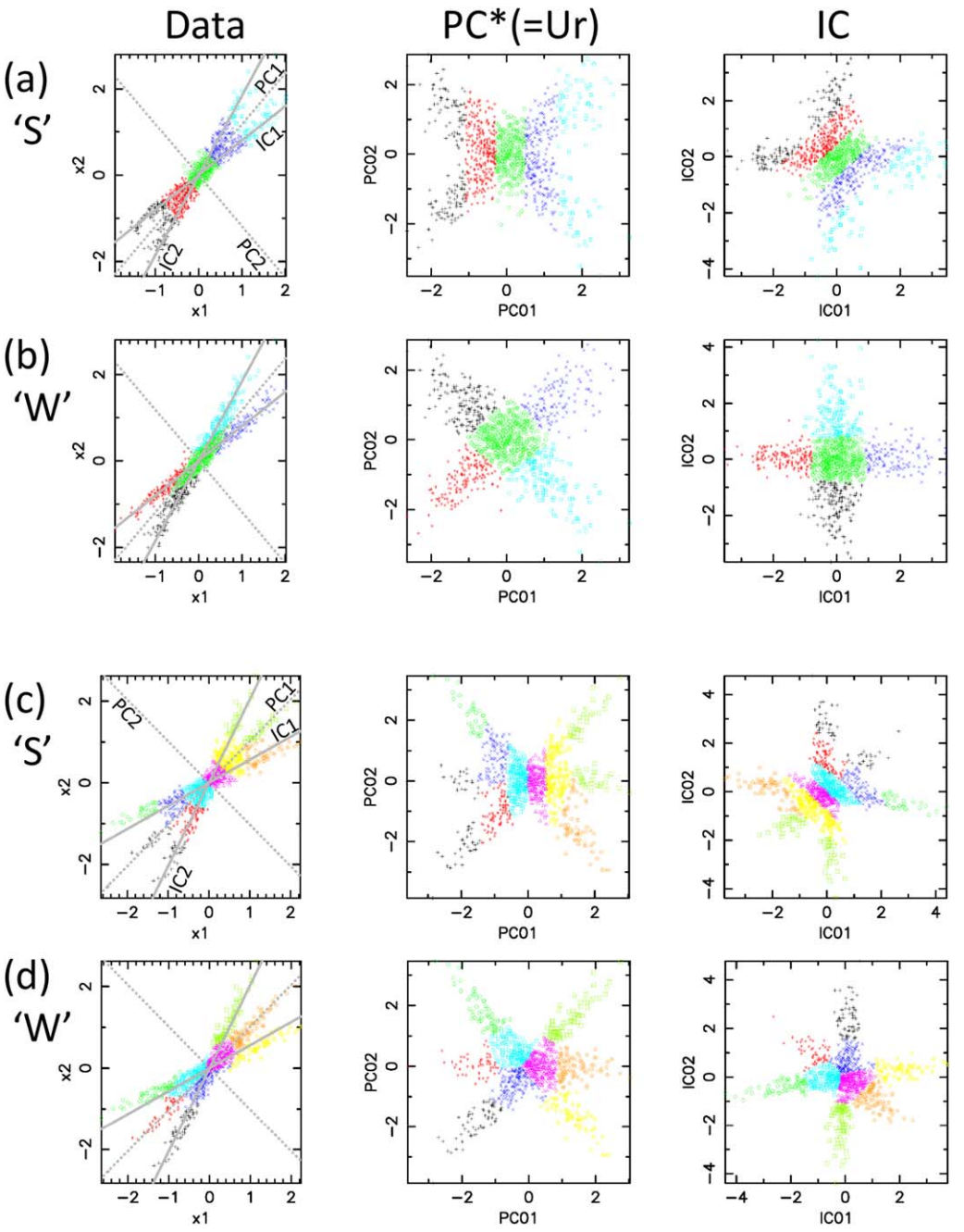

Figure 5. Analyses of two sets of synthetic data [(a) and (b) with 1000 data points and (c) and (d) with 999 data points] with two variables. In Figures 5a and 5b, the left column plots the raw data with PCs and ICs as in "Fig. L" of Figure 2; the middle column shows the whitened data **U,** (=PC*, normalized PC scores) as in "Fig. M" of Figure 2; and the right column plots the IC scores as in "Fig. R" of Figure 2. The cluster number is assigned to be $K = 5$ for Figures 5a and 5b, and $K = 9$ for Figures 5c and 5d, as shown by the color coding. In Figure 5a, the same data as in Figure 4(c) are used. Accordingly, the PCs and ICs are the same. The difference between Figures 5a and 5b is the preprocessing of "S" (primary standardization) in Figure 5a and "W" (whitening) in Figure 5b for KCA. In Figure 5c, a two-variable joint Gaussian distribution (with the zero mean and the standard deviations of unity and 0.1) is rotated 25°, 45°, and 65°, respectively, producing three crossing trends. The difference between Figures 5c and 5d is the preprocessing of "S" (primary standardization) in Figure 5c and "W" (whitening) in Figure 5d for KCA.

Sensitivity for PCA, ICA, and KCA by progressively undersampling the distributions is interesting and practically important. In supporting information Figures S1–S3, the PCA-ICA-KCA results that are comparable to Figure 4 but with fewer numbers of samples n (i.e., $n = 400$, 200, 100, and 50 compared to $n = 1000$ in Figure 4) are shown. The results on the whitened data show that, although several hundred or more sets of

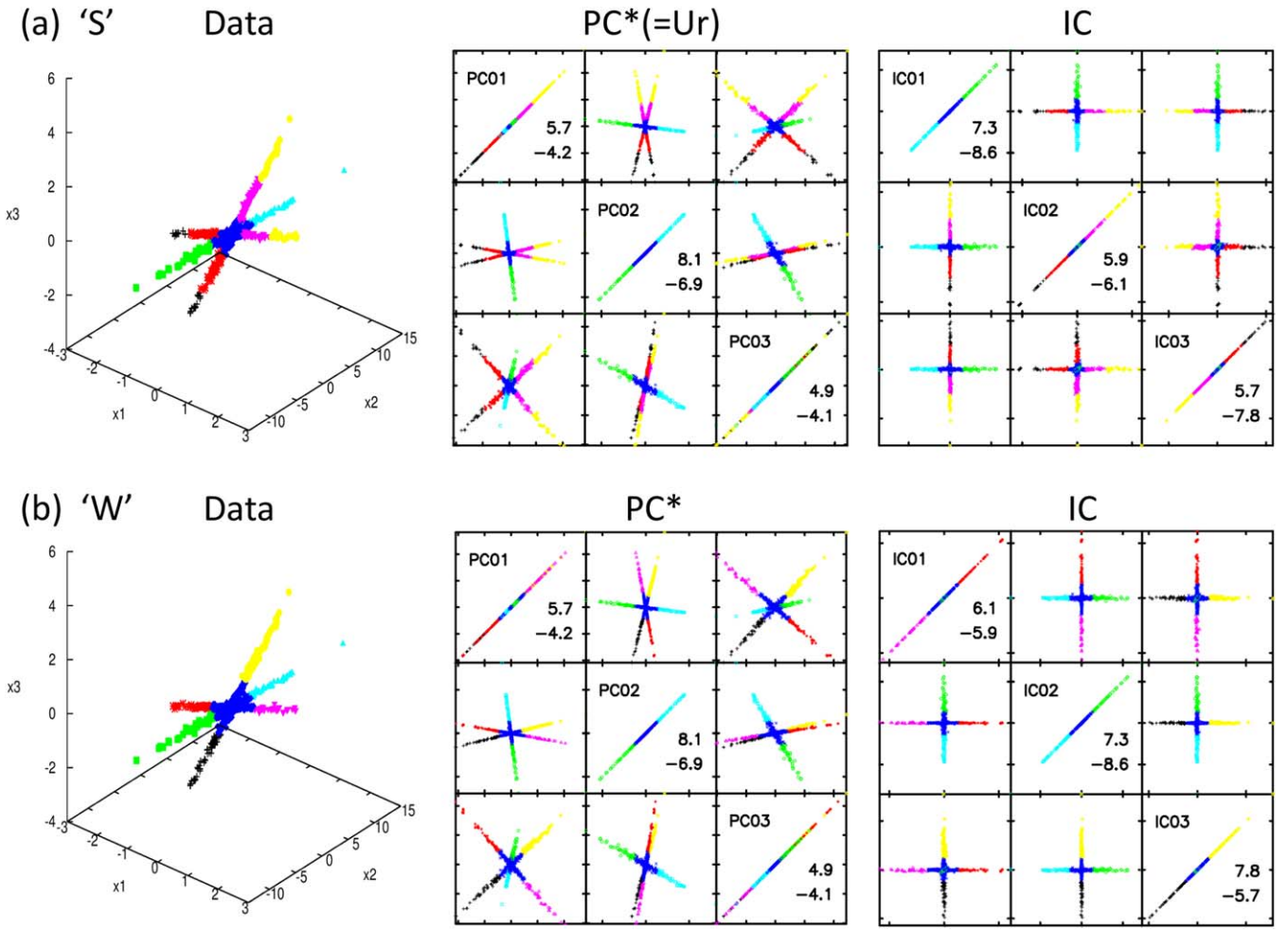

Figure 6. Analyses of a set of synthetic 999 data points with three variables. A three-variable joint Gaussian distribution (with a zero mean and the standard deviations of 2.0, 0.2, and 0.2 for (x, y, z), respectively) is transformed to $(x', y', z') = \left(\frac{x+y}{2}, \frac{y-x}{2}, \ 4(x+z)\right)$. Then, (x', y', z') is rotated by a certain angle a around first the x axis and second the y axis: $a = 15°$, 45°, and 90° are applied to produce three crossing trends. For clarity, the PC and IC vectors are not displayed in the left column. In the middle and right columns, the 2-D projections for PC*s and ICs are shown, where the numbers in the diagonal plots indicate the minimum and maximum of each PC or IC score/axis. The preprocessing of "S" (primary standardization) is applied to Figure 6a, whereas "W" (whitening) is applied to Figure 6b.

data are desirable for accurate analyses, the cases with even 50 may broadly capture the structure for two-dimensional data. Therefore, the methods proposed in this study may be applicable to a relatively small data set, although the number required for accurate analysis depends on the dimension and structure of data.

4. Discussion

4.1. Difference and Affinity of Methods

The results of KCA show remarkable differences in the clustering behavior depending on the type of data preprocessing. Primary standardization is a common method to cope with multivariate data consisting of different variables and units [*Varmuza and Filzmoser*, 2009]. With this type of preprocessing, the data may retain the original structure in terms of the variances, e.g., the elongation in a certain direction as in the left column of Figures 2–4. This direction coincides with that expressed by PC1, because PC1 is along the direction that maximizes the sample variance. Accordingly, the clusters are defined along PC1, irrespective of the number of clusters (K) assigned by users in Figure 3a and Figure 4a. If K is further increased, additional partitioning of the data along the subsequent PCs (i.e., PC2, PC3, . . .) will appear. Therefore, there is a direct link between KCA and PCA for the standardized data. This tight relationship between the two statistical

methods has been proved mathematically: The PCs are the continuous solutions to the discrete cluster membership indicators for KCA [*Ding and He*, 2004].

However, this is not the case when KCA is performed on the whitened data, as has been demonstrated in this study. The subspace consisting of the cluster centroids is identical to the ICA subspace spanned by the IC directions or divided by the IC axes that also bounds the clusters in some cases (Figures 4b, 4d, and 4f). Therefore, a tight link exists between KCA and ICA for the whitened data.

It is noted that if the data are expressed using PC scores and KCA is performed, KCA gives the same solution to that given by the primary standardization. However, if the PC scores consisting of m components are normalized by the corresponding eigenvalues, i.e., PC*, the middle columns in Figures 2–4, the distances between the data points are the same as those of the whitened data points, which gives the same clustering results as those in (b), (d), and (f) of Figures 2–4.

If the PC* data set (e.g., middle column of Figure 4d) is rotated, it coincides with the data in the IC space spanned by ICs (right column of Figure 4d), when the non-Gaussianity of marginal probability distribution is maximized instead of variance, as shown in Figure 7 of *Iwamori and Nakamura* [2015]. Accordingly, the PCs reflect, in a sense, amplitude or power with which the samples have been produced, whereas the ICs reflect independent structure or uniqueness associated with the original sources and process irrespective of the amplitude.

4.2. Advantages-Disadvantages and Complementary Roles of the Methods

The differences between PCs and ICs explain why PCs, particularly PC1, are oblique to the independent trends or group alignment in Figures 2–4, whereas ICs successfully identify them. This is always the case with data exhibiting multivariate non-Gaussian distribution for which PCA fails to identify the independent features. If the independent processes and sources are to be identified, ICA must be used. The discretized solution of ICA may be obtained from the KCA of the whitened data, which is in some cases more convenient for comparison with the discretized observations such as rock types or tectonic setting, as will be shown later. Once the independent elementary processes or sources are recovered, the mechanism by which the data distribution is created can be deduced (e.g., how the independent processes are coupled or mixed to produce the apparent overall variance). At the same time, the overall variation and its amplitude are soundly and economically determined by computation with PCA, which gives a continuous solution, and KCA, which provides a discretized solution, based on primary standardization. Quantifying the amplitude of variation by PCA is important, partly because ICA cannot determine the amplitude associated with each IC. Therefore, the approach combining KCA, PCA, and ICA will provide a powerful method for looking into the data structure and its origin.

In any case, primary standardization, conducted by normalizing the raw data with the mean and the variance, and PCA, or eigenvalues and eigenvectors of the correlation coefficient matrix, are fundamental to these methods. Then, ICA can be a powerful tool for understanding the independent structures, which feeds back to the interpretation of PCA. However, ICA is computationally consuming, and the solution is generally nonunique because multiple sets of solutions could be obtained depending on the initial condition of the solution search procedure. This also requires time and effort for the interpretation. Considering these advantages and disadvantages of KCA, PCA, and ICA, an economical procedure is to perform KCA on the whitened data, which requires first PCA and gives a similar solution to ICA. Depending on the number of clusters assigned, problems with the computational time and nonuniqueness of the solution could arise. Several advanced methods can be used to avoid these problems, such as k-means ++ [*Arthur*, 2007]; however, a trial with a feasible number of initial conditions may converge the solution rather well, and it is worth trying with the included Excel program "KCA."

Estimation of the optimal number of clusters is a still major challenge in KCA [*Tibshirani et al*., 2001; *Sugar and James*, 2003; *Jain*, 2010]. Numerous approaches to this problem have been suggested over the years, including Calinski and Harabasz's index, Hartigan's rule, the Kranowski and Lai test, the silhouette statistic, a Gaussian model-based approach, the gap statistic, and an information theoretic approach [*Sugar and James*, 2003, and the references therein]. Unfortunately, no perfect mathematical criterion is available for choosing K. Typically, KCA is run independently for different values of K and the partition that appears the most meaningful to the domain expert is selected [*Jain*, 2010]. As will be shown in section 4.3, multiple numbers of clusters should be examined depending on the individual problems.

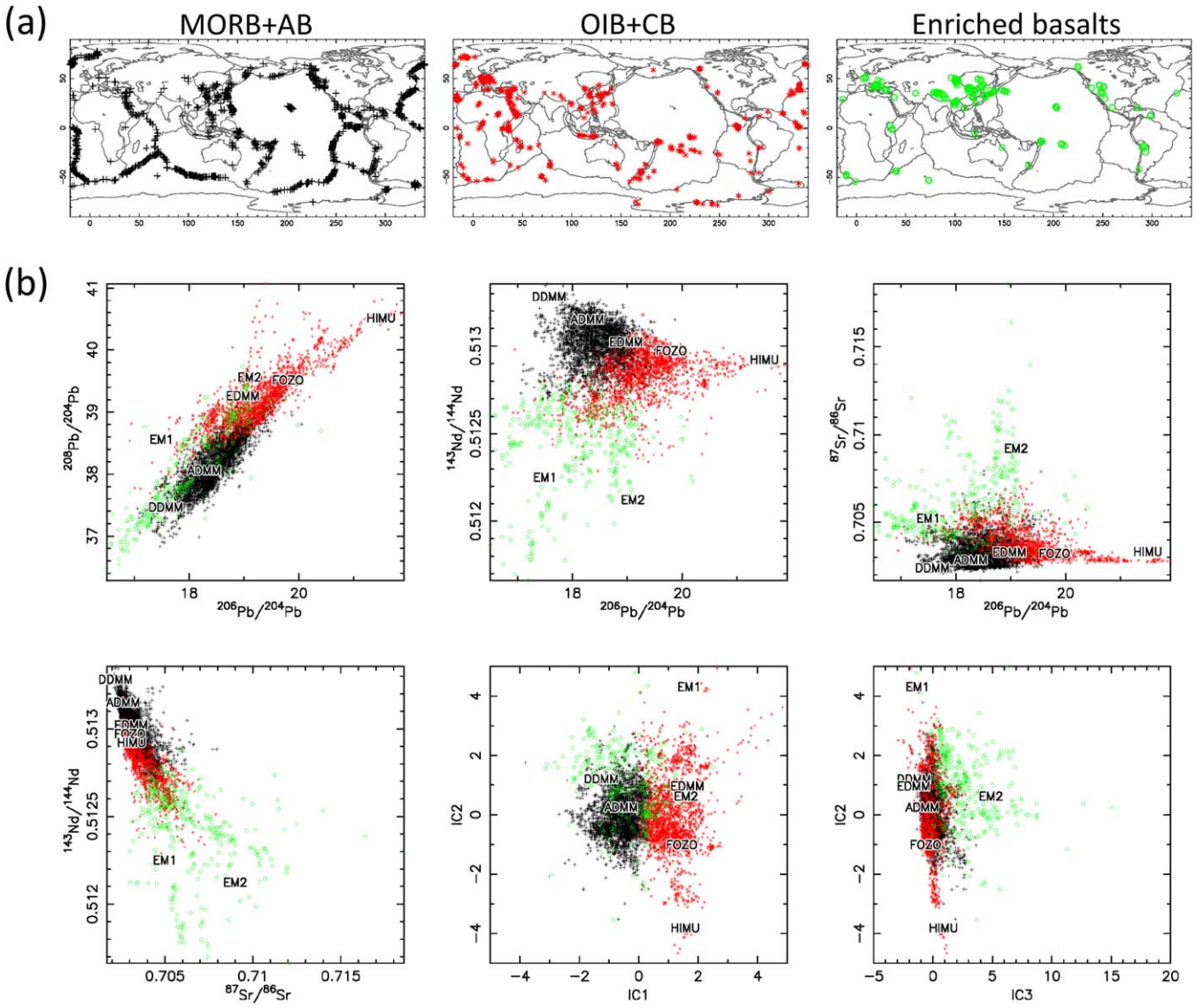

Figure 7. Results of KCA for 6854 samples with five isotopic ratios ($^{87}Sr/^{86}Sr$, $^{143}Nd/^{144}Nd$, $^{206}Pb/^{204}Pb$, $^{207}Pb/^{204}Pb$, and $^{208}Pb/^{204}Pb$) from mid-ocean ridges, ocean islands, arcs, and continental areas [*Iwamori and Nakamura*, 2015]. The number of clusters $K=3$ and the whitening preprocessing of data have been applied to obtain the clusters shown in (a) the geographical maps and (b) the compositional diagrams. In the ascending order of $^{87}Sr/^{86}Sr$ of the centroids, the three clusters broadly correspond to (i) mid-ocean ridge basalt + arc basalt (MORB+AB), (ii) ocean island basalt + continental basalt (OIB + CB), and (iii) enriched basalts including EM1 and EM2 groups. As reported by *Iwamori and Nakamura* [2012, 2015], some of the AB are classified into (ii), and the two representative hotspots, Hawaii and Iceland, are classified into (i). In Figure 7b, the representative mantle geochemical end-members are shown after *Iwamori and Nakamura* [2015]: DMM (Depleted MORB Mantle), HIMU (high-μ basalt), FOZO/C (FOcus ZOne/Common component), EM1 (Enriched Mantle 1), EM2 (Enriched Mantle 2).

Nevertheless, some broad criteria for choosing K are suggested on the basis of synthetic data analysis as was discussed for Figures 3–5. The number of clusters K should be greater than the reduced data dimension r, which as discussed in section 2 can be determined on the basis of eigenvalues for the correlation coefficient matrix. Increasing K from r will increase the resolution for the main part where the number density of data is high within the compositional space, as well as specific trends that branch from the main part (Figures 5b and 5d), which may avoid artificial segmentation. Because at least two clusters and their centroids are required to identify a compositional vector for each dimension, including possible branches, we suggest $K=r$ to $3r$ as a reasonable range to search for an appropriate K. If we use too many clusters, finding the positions and directions of clusters in the data structure becomes difficult [*Tan et al.*, 2004], and the partition could approach an extreme case with $K=$ the number of samples, which is meaningless. By compromising these competing factors, K is determined by trial and error.

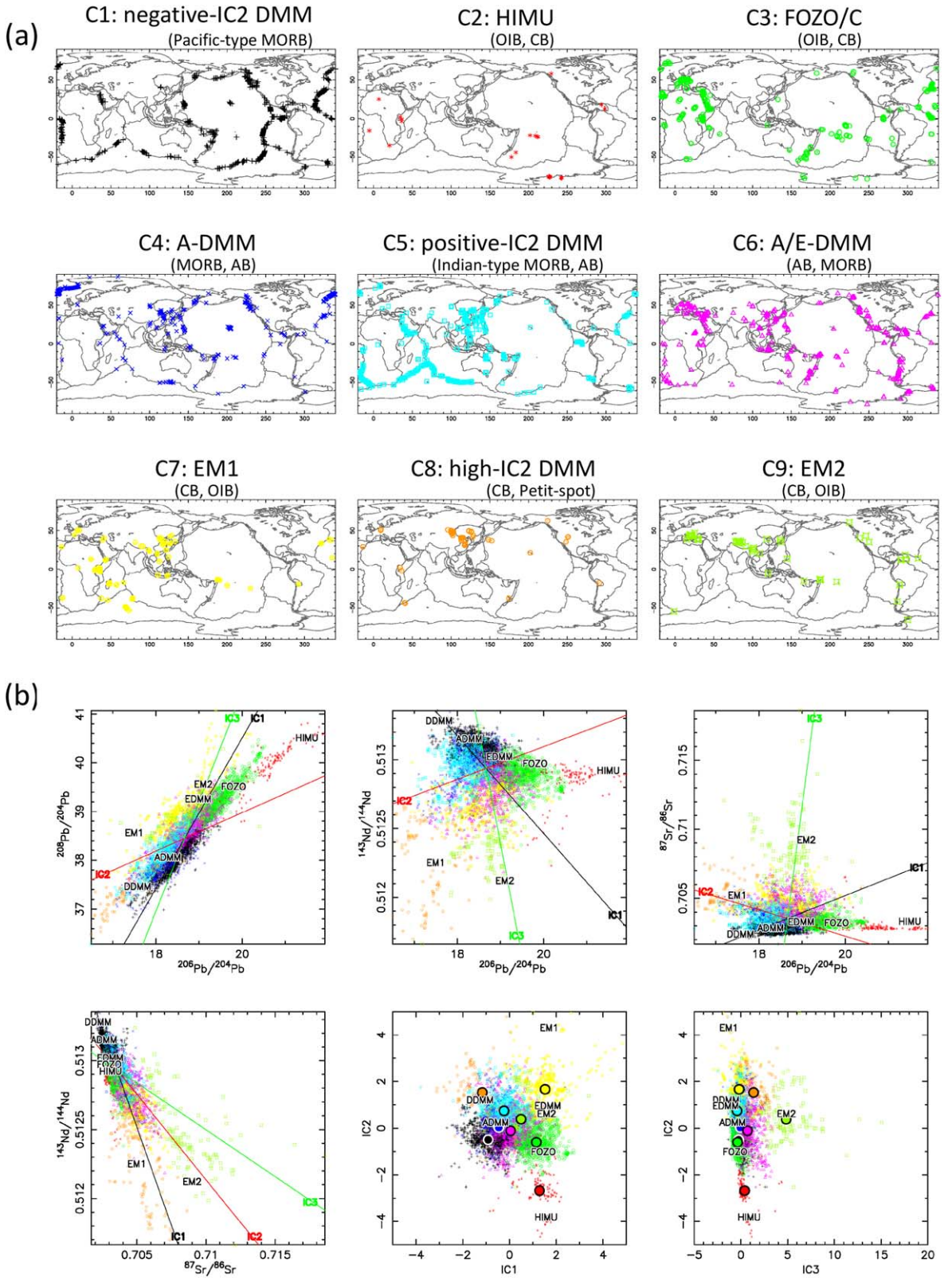
(a)
C1: negative-IC2 DMM
(Pacific-type MORB)
C2: HIMU
(OIB, CB)
C3: FOZO/C
(OIB, CB)
C4: A-DMM
(MORB, AB)
C5: positive-IC2 DMM
(Indian-type MORB, AB)
C6: A/E-DMM
(AB, MORB)
C7: EM1
(CB, OIB)
C8: high-IC2 DMM
(CB, Petit-spot)
C9: EM2
(CB, OIB)
(b)
$^{208}Pb/^{204}Pb$
$^{206}Pb/^{204}Pb$
$^{143}Nd/^{144}Nd$
$^{87}Sr/^{86}Sr$
IC1
IC2
IC3
DDMM
ADMM
EDMM
FOZO
HIMU
EM1
EM2

Figure 8.

4.3. Application to Isotopic Composition of Basalt

Full application of these methods will be presented elsewhere. Here an example for KCA on whitened data is shown, on the basis of the isotopic compositions of young basalts from almost all tectonic settings around the globe [*Iwamori and Nakamura*, 2015]. The data consist of 6854 samples with five isotopic ratios ($^{87}Sr/^{86}Sr$, $^{143}Nd/^{144}Nd$, $^{206}Pb/^{204}Pb$, $^{207}Pb/^{204}Pb$, and $^{208}Pb/^{204}Pb$) from mid-ocean ridges (MORB), ocean islands (OIB), arcs (AB), and continental areas (CB). A series of detailed inspections using ICA has been performed on the data [*Iwamori et al*., 2010; *Iwamori and Nakamura*, 2012, 2015], and the results of KCA were compared with the previous ICA results. The KCA result is a solution in a discrete format; therefore, comparison of the clusters and the basalt types and tectonic settings including MORB, OIB, AB, CB, and mantle geochemical end-members such as DMM and EM, which also have discretized formats, may be easier than the continuous solution of ICA.

Because the three PCs account for more than 98% of the sample variance, first, the number of clusters, $K=3$, is assigned for KCA on the whitened data. The three clusters broadly correspond to "MORB + AB," "OIB + CB," and "Enriched basalts," respectively (Figure 7a). The first cluster shows the most depleted signatures, as represented by the depleted MORB mantle (DMM) composition (Figure 7b). The second cluster has a mean value that is distinctly high in $^{206}Pb/^{204}Pb$, $^{207}Pb/^{204}Pb$, and $^{208}Pb/^{204}Pb$, and low in $^{143}Nd/^{144}Nd$ but nearly the same $^{87}Sr/^{86}Sr$ compared with the first cluster. These discriminations correspond to IC1 in ICA performed by *Iwamori and Nakamura* [2015], as shown in the IC1-IC2 diagram (Figure 7b): "MORB + AB" correspond to negative IC1, and "OIB + CB" correspond to positive IC1. Accordingly, IC1 broadly discriminates OIB from MORB. Except for Hawaii and Iceland, 95% of OIB plotted on the positive IC1 field. Moreover, 83% of MORB plot on the negative IC1 field except for those from plume-influenced ridges such as Iceland, Azores, Galapagos, and the Red Sea [*Iwamori and Nakamura*, 2015]. The third cluster shows high $^{87}Sr/^{86}Sr$ and low $^{143}Nd/^{144}Nd$, corresponding to basalts mostly from OIB + CB. This cluster involves EM1 and EM2 basalts with variability in both the IC2 and IC3 directions of *Iwamori and Nakamura* [2015], as shown in the IC3-IC2 diagram in Figure 7b. Along the IC2 direction, the basalts from the Eastern and Western Hemispheres have been discriminated [*Iwamori and Nakamura*, 2012, 2015], which is not seen in Figure 7b. Therefore, as demonstrated by Figures 2–4, the number of clusters, $K=3$ in this case, is insufficient to fully resolve the IC space, particularly for the IC2 and IC3 directions.

Figure 8 shows the result of KCA when $K=9$, which can be regarded as a finely clustered version of Figure 7. This result exhibits broad geographical and tectonic provenance as follows: "MORB + AB" of Figure 7 is approximately divided into C1, C4, C5, and C6; "OIB + CB" is divided into C2 and C3; and "Enriched basalts" into C7, C8, and C9, respectively (Figure 8a). The nine clusters, C1 to C9, broadly correspond to the following mantle geochemical end-members and IC configurations (Figure 8b): C1: negative-IC2 DMM [IC1−, IC2−]; C2: HIMU [IC1+, IC2−−]; C3: FOZO/C [IC1+, IC2−]; C4: A-DMM [IC1−, IC2∼0]; C5: positive-IC2 DMM [IC1−, IC2+]; C6: transitional range between A-DMM and E-DMM [IC1∼0 to +, IC2∼0 to +]; C7: EM1 [IC1+ to ++, IC2+ to ++]; C8: high-IC2 DMM [IC1−−, IC2++]; and C9: EM2 [IC3++], where $++ > + > 0 > - > --$ for the IC scores in the IC1-IC2 and IC3-IC2 diagrams (Figure 8b).

These clusters and IC values reflect the geochemical nature of source materials created by differentiation and homogenization processes. For the Rb-Sr, Sm-Nd, and U-Th-Pb systems shown in Figure 8, the compositional vectors labeled as IC1 and IC2 have been argued to represent parent/daughter fractionation associated with melting and aqueous fluid-mineral reaction, respectively [*Iwamori and Albarède*, 2008; *Iwamori et al*., 2010; *Iwamori and Nakamura*, 2015]. On the basis of the statistical model of *Rudge et al*. [2005] for the elemental partitioning and subsequent radiogenic ingrowth, the fractionation vector, particularly its slope, was calculated and compared with the IC vectors. In this case, positive/negative IC1 values indicate long-

Figure 8. Results of KCA for 6854 samples with five isotopic ratios ($^{87}Sr/^{86}Sr$, $^{143}Nd/^{144}Nd$, $^{206}Pb/^{204}Pb$, $^{207}Pb/^{204}Pb$, and $^{208}Pb/^{204}Pb$) from MORB, OIB, AB, and CB [*Iwamori and Nakamura*, 2015]. The number of clusters $K=9$ and the whitening preprocessing of data have been applied to obtain the clusters shown in (a) the geographical maps and (b) the compositional diagrams. In the ascending order of $^{87}Sr/^{86}Sr$ of the centroids, the nine clusters broadly correspond to specific mantle geochemical end-members and IC configurations: (C1) D-DMM (Pacific Ocean-type MORB), (C2) HIMU, (C3) FOZO/C, (C4) A-DMM, (C5) Moderately positive-IC2 DMM (Indian Ocean-type MORB), (C6) Arc basalt and the transitional range between A-DMM and E-DMM, (C7) EM1, (C8) High-IC2 DMM, and (C9) EM2. In Figure 8b, the representative mantle geochemical end-members are shown after *Iwamori and Nakamura* [2015]: DMM (Depleted MORB Mantle), HIMU (high-μ basalt), FOZO/C (FOcus ZOne/Common component), EM1 (Enriched Mantle 1), EM2 (Enriched Mantle 2). The three independent components (i.e., linear lines labeled as IC1, IC2, and IC3 with specific slopes in each diagram of Figure 8b) are plotted following *Iwamori and Nakamura* [2015]. In each diagram, labels IC1, IC2, and IC3 are placed along the positive sides of individual IC axes.

term melt component-enriched/depleted materials, corresponding to OIB/MORB sources, whereas positive/negative IC2 values indicate long-term aqueous fluid component-enriched/depleted materials, dividing the Earth's mantle into the Eastern/Western Hemispheres, respectively [*Iwamori and Nakamura*, 2012, 2015].

Because the 2-D plane spanned by IC1 and IC2 account for 95% of the sample variance in the original variable space [*Iwamori and Nakamura*, 2015], most of the data except for C9 are along the IC1-IC2 plane, suggesting that the main geochemical features of the mantle can be attributed to only two differentiation processes above. In addition to the separation of C9, which is the high-IC3 cluster, the E-W hemispherical structure is shown in C1 (including Pacific-type MORB that characterizes the Western Hemisphere) versus C5 (including Indian-type MORB) and C7 (including EM1 basalts) which characterize the Eastern Hemisphere. These clusters and discriminations represent the compositional domains in the IC space with sufficient resolution, with the following additional clusters for the compositional volume in which the number density of samples is high: DMM is subdivided into C1 [IC1−, IC2−], C4 [IC1−, IC2 ∼ 0], C5 [IC1−, IC2+], C6 [IC1 ∼ 0 to +, IC2 ∼ 0 to +], and C8 [IC1−−, IC2++]. The average composition of DMM (A-DMM) [*Workman and Hart*, 2005] roughly corresponds to C4. In the geographical distribution, it is not as prevalent as the other types of DMM (e.g., C1 and C5) and FOZO/C as a common source for OIB [*Hart et al*., 1992; *Hanan and Graham*, 1996]. High-IC2 DMM, represented by the Petit-spot basalts in the northwestern Pacific, corresponds to C8, a negative-IC1 and positive-IC2 domain in the IC space [*Iwamori et al*., 2010; *Iwamori and Nakamura*, 2015], which cannot be expressed by any linear combination of the mantle geochemical end-members proposed thus far. The centroids of individual clusters in the IC1-IC2 and IC3-IC2 diagrams of Figure 8b (colored solid circles) clearly show the quadratic partition of data on the major IC1-IC2 plane by four clusters (C1, C3, C7, and C8). In addition, the central part within the quadratic data distribution, where the number density of data is high, is divided into C4, C5, and C6. Furthermore, two clusters outside of the quadrant, C2 and C9, are discriminated: C2 is characterized by an extreme combination of high IC1 and low IC2, involving HIMU, and C9 is characterized by high IC3, involving EM2. In summary, the clusters found in this analysis are consistent with the previous ICA results and even clarify the data structure and how the compositional domains are to be defined beyond the combination and description with conventional mantle geochemical end-members.

The results between $K=4$ and $K=8$ show a transition between Figures 7 and 8, increasing the resolution for MORB-suites (e.g., C1, C4, and C5 for $K=9$) and OIB-suites (e.g., C2, C3, and C7) with increasing K (supporting information Figures S4 to S6 for $K=6,\ 7,$ and 8). Of course, assigning K divides the natural system artificially, and because the basalt types have significant overlap in compositions in nature, the discrimination is not well defined. However, the overall feature is soundly captured. Figures 7 and 8 show the potential usefulness of the multivariate methods, particularly for the high-dimensional data $(m \gg 3)$. Finally, we compared the results of KCA on the standardized data without whitening and the whitened data (supporting information Figures S7 (for $K=3$) and S8 ($K=9$)). In supporting information Figure S7, the separation of the data in terms of positive/negative IC1 (hence MORB/OIB discrimination) is not clear in the simple standardized KCA (supporting information Figure S7a), whereas the whitened data-based KCA separates positive/negative IC1 fields sharply (supporting information Figure S7b; Figure 7b). In supporting information Figure S8, the centroids of the nine clusters are arranged in a rather sheared manner for the simple standardized KCA, which indicates that the orthogonality of the base vectors is invalidated (supporting information Figure S8a), whereas the whitened data-based KCA represents the four quadrants of the IC1-IC2 space rather clearly (supporting information Figure S8b). These comparisons demonstrate the advantage of the new technique presented in this study.

5. Conclusions

The important results in this study are summarized in Figure 2, and three fundamental methods of multivariate statistical analysis, k-means cluster analysis (KCA), principal component analysis (PCA), and independent component analysis (ICA) have been tested for identifying the data structures including trends and clusters. Although KCA returns a discrete set of solutions including cluster indices for each sample and centroid for each cluster, and PCA and ICA provide a set of base vectors that describe the sample variability in the continuous space, they are found to have affinities. KCA and PCA for standardized data are useful for identifying the data structure controlled by the amplitude of variability, whereas KCA and ICA for whitened data are

able to identify the independent structures hidden in the data. Each method has unique advantages and disadvantages. PCA is a common and robust method, and may work as a preprocessing tool for whitening, whereas it fails to identify the independent structures for joint non-Gaussian distributions that occur in many natural systems. ICA is capable of identifying the independent structure; however, its solution may involve nonuniqueness, requiring careful interpretation. KCA may provide essentially the same information as either PCA or ICA, depending on the type of preprocessing, i.e., standardization or whitening, and is convenient when discretized classification is compared. In summary, the combination of all three methods may provide a complete view on the data structure. Of the three methods and their combination, sufficient information may be obtained from KCA on the whitened data obtained from PCA as a preprocessing procedure. With an accumulation of multidimensional data and an increase in computer power, the hidden data structure will be further unraveled, for which the methods in this study will be relatively simple yet powerful. Various types of geochemical data can be analyzed with the methods presented in this paper, for which the Excel program "KCA" is provided (supporting information and http://dsap.jamstec.go.jp/).

Appendix : Formulation

First, primary standardization of the raw data is applied to avoid effects of differences in the units and ranges of individual variables on the statistical results as follows. Let $\mathbf{X}=(x_{ij})$ be the data matrix for the ith sample $(i=1,2,\ldots,n)$ and the jth element/isotope j. The raw data are first standardized as

$$\mathbf{X}'=x'_{ij}=\left(x_{ij}-\bar{x}_j\right)/\sigma_j \tag{A1}$$

where the mean is $\bar{x}_j=\sum_{i=1}^{n} x_{ij}/n$ and the standard deviation is $\sigma_j=\sqrt{\frac{1}{n-1}\sum_{i=1}^{n}\left(x_{ij}-\bar{x}_j\right)^2}$. Several other preprocessing procedures can be used as described in section 2.

Then, the standardized (or preprocessed) data can be analyzed by KCA and/or PCA. KCA is a classification method widely used to partition the multivariate data such that the total distance between the mean of a cluster (centroid) and the individual data points in the cluster is minimized [*MacQueen*, 1967; *Jain*, 2010]. The KCA method partitions the data into a set of K clusters, $\boldsymbol{C}=\{c_k\}_{k=1,\ldots,K}$, that are determined by minimizing the sum of squared distance:

$$J_K(\boldsymbol{C})=\sum_{k=1}^{K}\sum_{i\in C_k}\sum_{j=1}^{m}\left(x_{ij}-\mu_{kj}\right)^2, \text{ and } \arg\min_C J_K(\boldsymbol{C}) \tag{A2}$$

where $\mu_{kj}=\sum_{i\in C_k} x_{ij}/n_k$ is the centroid of cluster c_k for jth element/isotope and n_k is the number of data in c_k. A standard iterative solution to KCA may be trapped in the local minima [*Bradley and Fayyad*, 1998]. Accordingly, a total of 100,000 trails was performed under randomly different initial conditions to find the global minimum in this study with $n=1000$ to $10,000$.

PCA is also a method commonly used to specify the uncorrelated base vectors that account for the data effectively by maximizing the variance along the principal components (PCs). The eigenvalue decomposition of $m\times m$ variance-covariance matrix for x'_{ij} (i.e., correlation coefficient matrix for x_{ij}) gives eigenvectors v_{jl} (i.e., principal component vectors or referred to as "factor loadings" of individual PCs) for the lth eigenvalues d_l $(l=1,2,\ldots,m)$. The PC scores for individual data points are given by

$$s_{il}=x'_{ij}\cdot v_{jl}. \tag{A3}$$

Then the whitened (or sphered) data u_{il} is obtained as

$$u_{il}=s_{il}/\sqrt{d_l}. \tag{A4}$$

The reduced whitened data $\mathbf{U}_r=(u_{il})_{l=1,2,\ldots,r}$, where $r\leq m$ depending on the degree of the data reduction, can be analyzed by KCA and ICA. The data reduction is commonly based on the eigenvalues, where the PC components with sufficiently large d_l are selected. This brings several advantages such as computation saving, noise removal, calculation stabilization; however, important information for KCA and ICA may be lost. In this study, the data are not reduced, keeping $r=m$, partly because $m=2$ or 3 only for the synthetic data in this study. The data whitening can be performed straight from the raw data by singular value decomposition. In this case, the eigenvalues and eigenvectors are affected significantly by the differences in range and

unit of the individual variables and are generally not suitable for geochemical data that involve such differences for common cases.

Compared to KCA and PCA, ICA is a less common but powerful method used to extract independent base vectors and has been employed to solve various problems in information and brain sciences [*Hyvärinen, et al*., 2001], including informatics of geochemical data [*Iwamori and Albarède*, 2008; *Iwamori et al*., 2010]. The independent components (ICs, or IC vectors) can be found by rotating the ICs to maximize the non-Gaussianity J_G as follows: (i) searching the ith independent component vector $\mathbf{w}_i=(w_{li})$ to maximize J_G :

$$J_G(\mathbf{y})=\{E[G(\mathbf{y})]-E[G(v)]\}^2, \text{ and } \arg\max_{\mathbf{y}} J_G(\mathbf{y}), \tag{A5}$$

where $G(a)=-\exp(-a^2/2)$, $\mathbf{y}=\mathbf{U}_r\,\mathbf{w}_i$ (IC scores for the individual data points), E represents expectation, and v returns a standardized Gaussian distribution, and (ii) finding another $\mathbf{w}_{i+1}$ in the space orthogonal to $\mathbf{w}_{1,2,\ldots,i}$ until the number of independent components reaches r [*Iwamori and Nakamura*, 2015].

Acknowledgments
The authors thank Shun-suke Horiuchi and Mika Mohamed Abdelbaky Seif El-Nasr for their support, and John Maclennan, Andreas Stracke, and an anonymous reviewer for their constructive comments. This work was supported by the JSPS KAKENHI grant 26247091 and 26109006 for H.I., 16K00531 for H.N., 15H05833 for K.U., and MEXT KAKENHI grant 25120005 for T.K. and K.Y.

References

Agrawal, S., M. Guevara, and S. P. Verma (2004), Discriminant analysis applied to establish major-element field boundaries for tectonic varieties of basic rocks, *Int. Geol. Rev.*, *46*, 575–594, doi:10.2747/0020-6814.46.7.575.
Aitchison, J. (1986), *The Statistical Analysis of Compositional Data*, Chapman and Hall, London.
Allègre, C. J., B. Hamelin, A. Provost, and B. Dupré (1987), Topology in isotopic multispace and origin of the mantle chemical heterogeneities, *Earth Planet. Sci. Lett.*, *81*, 319–337.
Arthur, D., (2007), K-means++: The advantages of careful seeding, in *Proceedings of the 18th Annual ACM-SIAM Symposium on Discrete Algorithm*, pp. 1027–1035, Society for Industrial and Applied Mathematics, New Orleans, Louisiana.
Bowen, N. L. (1928), *The Evolution of the Igneous Rocks*, Princeton, N. J.
Bradley, P., and U. Fayyad (1998), Refining initial points for k-means clustering, in *Proceedings of the 15th International Conference on Machine Learning*, Morgan Kaufmann Publishers Inc., Burlington, Mass.
Brandmeier, M., and G. Wörner (2016), Compositional variations of ignimbrite magmas in the Central Andes over the past 26 Ma—A multivariate statistical perspective, *Lithos*, *262*, 713–728.
Ding, C., and X. He (2004), K-means clustering via principal component analysis, in *Proceedings of the 21st International Conference on Machine Learning*, Banff, Canada, Association for Computing Machinery, New York.
Fenner, C. N. (1929), The crystallizations of basalts, *Am J. Sci. 5th Ser.*, *18*, 225–253.
Hamilton, W. C (1964), *Statistics in Physical Science*, George Ronald, New York.
Hanan, B. B., and D. W. Graham (1996), Lead and helium isotope evidence from oceanic basalts for a common deep source of mantle plumes, *Science*, *272*, 991–995.
Hart, S. R., E. H. Hauri, L. A. Oschmann, and J. A. Whitehead (1992), Mantle plumes and entrainment: Isotopic evidence, *Science*, *256*, 517–520.
Hickey-Vargas, R. (1998), Origin of the Indian Ocean-type isotopic signature in basalts from Philippine Sea plate spreading centers: An assessment of local versus large-scale processes, *J. Geophys. Res.*, *103*, 20,963–20,979.
Hofmann, A. W. (2003), Sampling mantle heterogeneity through oceanic basalts: Isotopes and trace elements, in *The Mantle and Core, Treatise on Geochemistry 2*, edited by R. W. Carlson, pp. 61–101, Elsevier, Amsterdam.
Hyvärinen, A. (1999), Fast and robust fixed-point algorithms for independent component analysis, *IEEE Trans. Neural Networks*, *10*, 626–634.
Hyvärinen, A., J. Karhunen, and E. Oja (2001), *Independent Component Analysis*, John Wiley, N. J.
Iwamori, H., and F. Albarède (2008), Decoupled isotopic record of ridge and subduction zone processes in oceanic basalts by independent component analysis, *Geochem. Geophys. Geosyst.*, *9*, Q04033, doi:10.1029/2007GC001753.
Iwamori, H., and H. Nakamura (2012), East-west geochemical hemispheres anchored to asthenosphere in the Earth's mantle, *Geochem. J.*, *46*, e39–e46, doi:10.2343/geochemj.2.0224.
Iwamori, H., and H. Nakamura (2015), Isotopic heterogeneity of oceanic, arc and continental basalts and its implications for mantle dynamics, *Gondwana Res.*, *27*, 1131–1152, doi:10.1016/j.gr.2014.09.003.
Iwamori, H., F. Albarède, and H. Nakamura (2010), Global structure of mantle isotopic heterogeneity and its implications for mantle differentiation and convection, *Earth Planet. Sci. Lett.*, *299*, 339–351.
Jain, A. K. (2010) Data clustering: 50 years beyond K-means, *Pattern Recognit. Lett.*, *31*, 651–666.
Jenner, F. E, and H. St. C. O'Neil (2012), Analysis of 60 elements in 616 ocean floor basaltic glasses, *Geochem. Geophys. Geosyst.*, *13*, Q02005, doi:10.1029/2011GC004009.
Kuwatani, T., K. Nagata, M. Okada, T. Watanabe, Y. Ogawa, T. Komai, and N. Tsuchiya (2014), Machine-learning techniques for geochemical discrimination of 2011 Tohoku tsunami deposits, *Sci. Rep.*, *4*, 7077, doi:10.1038/srep07077.
Le Maitre, R. W. (1982), Numerical petrology, in *Statistical Interpretation of Geochemical Data*, vol. 8, *Developments in Petrology*, Elsevier, Amsterdam.
Lehnert, K., Y. Su, C. Langmuir, B. Sarbas, and U. Nohl (2000), A global geochemical database structure for rocks, *Geochem. Geophys. Geosyst.*, *1*, 1012, doi:10.029/1999GC000026.
MacQueen, J. (1967), Some methods for classification and analysis of multivariate observations, in *Proceedings of the Fifth Berkeley Symposium on Mathematical Statistics and Probability*, pp. 281–297, Univ. of Calif. Press, Calif.
Miyashiro, A. (1975), Volcanic rock series and tectonic setting, *Annu. Rev. Earth Planet. Sci.*, *3*, 251.
Nakamura, H., and H. Iwamori (2013), Generation of adakites in a cold subduction zone due to double subducting plates, *Contrib. Mineral. Petrol.*, *165*, 1107–1134, doi:10.1007/s00410-013-0850-0.
Petrelli, M, and D. Perugini, (2016), Solving petrological problems through machine learning: The study case of tectonic discrimination using geochemical and isotopic data, *Contrib. Mineral. Petrol.*, *172*, 81, doi:10.1007/s00410-016-1292-2.
Pisias, N. G., R. W. Murray, and R. P. Scudder (2013), Multivariate statistical analysis and partitioning of sedimentary geochemical data sets: General principles and specific MATLAB scripts, *Geochem. Geophys. Geosyst.*, *14*, 4015–4020, doi:10.1002/ggge.20247.

Reimann, C., P. Filzmoser, R. Garrett, and R. Dutter (2011), *Statistical Data Analysis Explained: Applied Environmental Statistics with R*, John Wiley, N. J.

Rudge, J. F., D. McKenzie, and P. H. Haynes (2005), A theoretical approach to understanding the isotopic heterogeneity of mid-ocean ridge basalt, *Geochim. Cosmochim. Acta*, *69*, 3873–3887.

Snow, C. A. (2006), A reevaluation of tectonic discrimination diagrams and a new probabilistic approach using large geochemical databases: Moving beyond binary and ternary plots, *J. Geophys. Res.*, *111*, B06206, doi:10.1029/2005JB003799.

Stracke, A. (2012), Earth's heterogeneous mantle: A product of convection-driven interaction between crust and mantle, *Chem. Geol.*, *330–331*, 274–299.

Sugar, C. A., and G. M. James (2003), Finding the number of clusters in a data set: An information theoretic approach, *J. Am. Stat. Assoc.*, *98*, 750–763, doi:10.1198/016214503000000666.

Tan, P.-N., M. Steinback, and V. Kumar (2004), *Introduction to Data Mining*, 736 pp., Addison-Wesley, Boston.

Tibshirani, R., G. Walther, and T. Hastie (2001), Estimating the number of clusters in a data set via the gap statistic, *J. R. Stat. Soc., Ser. B*, *63*, 411–423.

Usui, T., and H. Iwamori (2013), Mixing relations of the howardite-eucrite-diogenite suite: A new statistical approach of independent component analysis for the Dawn mission, *Meteorit. Planet. Sci.*, *48*, 2289–2299, doi:10.1111/maps.12205.

Varmuza, K., and P. Filzmoser (2009), *Introduction to Multivariate Statistical Analysis in Chemometrics*, Melbourne, CRC Press, New York.

Verma, S. P., K. Pandarinath, S. K. Verma, and S. Agrawal (2013), Fifteen new discriminant-function-based multi-dimensional robust diagrams for acid rocks and their application to Precambrian rocks, *Lithos*, *168–169*, 113–123. doi:10.1016/j.lithos.2013.01.014.

Vermeesch, P. (2006), Tectonic discrimination diagrams revisited, *Geochem. Geophys. Geosyst.*, *7*, Q06017, doi:10.1029/2005GC001092.

White, W. M., and R. A. Duncan (1996), Geochemistry and geochronology of the Society Islands: New evidence for deep mantle recycling, *Geophys. Monogr.*, *95*, 183–206.

Workman, R. K., and S. R. Hart (2005), Major and trace element composition of the depleted MORB mantle (DMM), *Earth Planet. Sci. Lett.*, *231*, 53–72.

Yasukawa, K., K. Nakamura, K. Fujinaga, H. Iwamori, and Y. Kato (2016), Tracking the spatiotemporal variations of statistically independent components involving enrichment of rare-earth elements in deep-sea sediments, *Sci. Rep.*, *6*, 29,603, doi:10.1038/srep29603.

Zindler, A., and S. R. Hart (1986), Chemical geodynamics, *Annu. Rev. Earth Planet. Sci.*, *14*, 493–571.

Zindler, A., E. Jagoutz and S. Goldstein (1982), Nd, Sr and Pb isotopic systematics in a three component mantle: A new perspective, *Nature*, *298*, 519–523.