Improved classi cation for compositional data using the α-transformation

Tsagris, Michail and Preston, Simon and T.A. Wood, Andrew

Department of computer science, University of Crete, Heraklion, Greece, School of Mathematical Sciences, University of Nottingham, UK, School of Mathematical Sciences, University of Nottingham, UK

2016

Improved classification for compositional data using the α-transformation

Michail Tsagris[1], Simon Preston[2] and Andrew T.A. Wood[2]

[1] Department of computer science, University of Crete, Heraklion, Greece

[2] School of Mathematical Sciences, University of Nottingham, UK

mtsagris@yahoo.gr, Simon.Preston@nottingham.ac.uk and Andrew.Wood@nottingham.ac.uk

Abstract

In compositional data analysis an observation is a vector containing non-negative values, only the relative sizes of which are considered to be of interest. Without loss of generality, a compositional vector can be taken to be a vector of proportions that sum to one. Data of this type arise in many areas including geology, archaeology, biology, economics and political science. In this paper we investigate methods for classification of compositional data. Our approach centres on the idea of using the α-transformation to transform the data and then to classify the transformed data via regularised discriminant analysis and the k-nearest neighbours algorithm. Using the α-transformation generalises two rival approaches in compositional data analysis, one (when α=1) that treats the data as though they were Euclidean, ignoring the compositional constraint, and another (when $\alpha = 0$) that employs Aitchison's centred log-ratio transformation. A numerical study with several real datasets shows that whether using $\alpha = 1$ or $\alpha = 0$ gives better classification performance depends on the dataset, and moreover that using an intermediate value of α can sometimes give better performance than using either 1 or 0.

Keywords: compositional data, classification, α-transformation, α-metric, Jensen-Shannon divergence

1 Introduction

Compositional data arise commonly in many fields, for instance geology (Aitchison, 1982), in studying constitution of rock samples; economics (Fry et al., 2000), in budget allocations; archaeology (Baxter et al., 2005), in the constitution of man-made glasses; and the political sciences (Rodriques and Lima, 2009), in voting behaviour. In compositional data analysis, a composition is considered an equivalence class comprising the set of multivariate vectors that differ only by a scalar factor and have non-negative components. Consequently, without loss of generality, an observation may be viewed as a vector of proportions, i.e., with non-negative components constrained to sum to 1. The sample space of the observations is hence the simplex

$$\mathbb{S}^d = \left\{ (x_1, ..., x_D)^T \middle| x_i \geq 0, \sum_{i=1}^{D} x_i = 1 \right\},$$

where D denotes the number of components of the vector and $d = D - 1$.

For statistical analysis of compositional data the question of how to account for the compositional constraint arises. A simple approach is to ignore the compositional constraint and treat the data as though they were Euclidean, an approach we will call "Euclidean data analysis"

(EDA) (Baxter, 2001; Baxter et al., 2005; Baxter and Freestone, 2006; Woronow, 1997). There is a school of thought, however, largely following from the work of Aitchison (1982, 1983, 1992), that ignoring the compositional constraint is inappropriate and can lead to misleading inferences. Aitchison contended that data should instead be analysed after applying a "logratio" transformation, arguing that this amounted to working with an implied distance measure on the simplex (discussed further in the next section) that satisfied particular mathematical properties he regarded as essential for compositional data analysis. Other approaches to compositional data analysis that we mention here but do not consider further in this paper include using different transformations, such as the square-root transformation (Stephens, 1982; Scealy and Welsh, 2011), and parametric modelling, for example using the Dirichlet distribution (Gueorguieva et al., 2008).

Both EDA and Aitchison's logratio analysis (LRA) approach are widely used and there has been a long and ongoing disagreement over which of these approaches, or indeed others, is most appropriate to use. The debate remains largely centred on the distance measures implied by the various approaches and whether or not they satisfy particular mathematical properties. Scealy and Welsh (2014) have recently presented a historical summary of the debate, and have given a critical appraisal of the properties often invoked by authors to support the use of LRA. We share Scealy and Welsh's opinion that LRA should not be a default choice for compositional data analysis on account of such properties. In this paper, we take the pragmatic view, which seems especially relevant for classification problems (in which out-of-sample classification error rate provides an objective measure of performance), that we should adopt whichever approach performs best in a given setting.

Indeed, a key message of this paper is that for classification problems, the choice of whether or not one should transform the data, and if so which transformation to use, should depend on the dataset under study. This conclusion is clear from the fact that we can easily generate a synthetic dataset for which LRA will perform perfectly and EDA poorly, and vice versa.

One characteristic of a dataset that immediately rules out using LRA in its standard form is the presence of observations for which one or more components is zero, since for such observations the logratio transformation is undefined. Data of this type are not uncommon (in §4 we consider two datasets containing observations with zeros), so this is a notable weakness of LRA. Some attempts have been made to modify LRA to make it appropriate for data containing zeros (particularly when the zeros are assumed to arise from rounding error), but these involve a somewhat ad hoc imputation approach of replacing zeros with small values. On a different tack, Butler and Glasbey (2008) developed parametric models specifically for compositional data with zeros.

Perhaps due to the nature of the data, little attention has been given to the problem of classifying compositional data, especially where zero values are present. Exceptions are Zadora et al. (2010) and Neocleous et al. (2011), who consider classification using parametric models to account for the possibility of zeros values; see also Palarea-Albaladejo et al. (2005) who consider the related problem of cluster analysis. Our goal in this paper is to develop adaptive classification algorithms which take into account the characteristics of individual datasets, such as the distribution of the groups and the presence of zeros. The main idea is to employ the Box–Cox-type α-transformation explored in Tsagris et al. (2011), and then use the transformed

data as a basis for classification. This transformation has a free parameter, α, and is such that the case $\alpha = 0$ corresponds to the logratio transformation, and $\alpha = 1$ corresponds to a linear transformation of the data. Hence using $\alpha = 0$ corresponds to LRA, and $\alpha = 1$, when used in conjunction with the discriminant analysis and nearest-neighbour classification algorithms that we consider in §3, is equivalent to EDA. For values of α between 0 and 1, the α-transformation offers a compromise between LRA and EDA. An important benefit of the α-transformation is that it is well-defined for any $\alpha > 0$ for compositions containing zeros.

The paper is structured as follows. In §2 we discuss in more detail the α-transformation and the logratio transformation, and the associated implied distance measures, and then in §3 we consider some classification techniques and how their performance can be improved using the α-transformation. In §4 we present the results of a numerical study with four real datasets to investigate the performance of the various techniques. We conclude in §5 with a discussion of the results.

2 The α-transformation and implied simplicial distance measure

The α-transformation of a compositional vector $\mathbf{x} \in \mathbb{S}^d$ (see Tsagris et al. (2011)) is defined by

$$\mathbf{z}_\alpha(\mathbf{x}) = \mathbf{H} \cdot \left(\frac{D\,\mathbf{u}_\alpha(\mathbf{x}) - \mathbf{1}_D}{\alpha} \right), \tag{1}$$

with $\alpha > 0$ (we discuss more general α below), and where

$$\mathbf{u}_\alpha(\mathbf{x}) = \left(\frac{x_1^\alpha}{\sum_{j=1}^D x_j^\alpha}, \ldots, \frac{x_D^\alpha}{\sum_{j=1}^D x_j^\alpha} \right)^T \tag{2}$$

is the compositional power transformation (Aitchison, 2003), $\mathbf{1}_D$ is the D-dimensional vector of ones, and $\mathbf{H}$ is any d-by-D matrix consisting of orthonormal rows, each of which is orthogonal to $\mathbf{1}_D$; similar ideas have been used in the compositional data context by Egozque et al. (2003) and, of course, in many other contexts. A suitable choice for $\mathbf{H}$ (noting in any case that the classification methods in this paper are invariant to the particular choice) is the Helmert matrix (Lancaster, 1965; Dryden and Mardia, 1998) with the first row removed, i.e., the matrix whose jth row is

$$(h_j, \ldots, h_j, -jh_j, 0, \ldots, 0), \quad \text{where} \quad h_j = -\left\{ j(j+1) \right\}^{-1/2}, \tag{3}$$

with h_j repeated j times and 0 repeated $d - j$ times. The purpose of $\mathbf{H}$ is to remove the redundant dimension which is present due to the compositional constraint. In particular, the vector $\left(D\,\mathbf{u}_\alpha(\mathbf{x}) - \mathbf{1}_D\right)/\alpha$ has components which sum to zero and therefore it lies in a subspace of $\mathbb{R}^D$; left-multiplication by $\mathbf{H}$ is an isometric one-to-one mapping from this subspace into $\mathbb{R}^d$. The image $\mathcal{V}_\alpha = \left\{ \mathbf{z}_\alpha(\mathbf{x}) : \mathbf{x} \in \mathbb{S}^d \right\}$ of transformation (1) is $\mathbb{R}^d$ in the limit $\alpha \to 0$ but a strict subset of $\mathbb{R}^d$ for $\alpha \neq 0$. Transformation (1) is invertible: for $\mathbf{v} \in \mathcal{V}_\alpha$ the inverse of $\mathbf{z}_\alpha(\mathbf{x})$ is

$$\mathbf{z}_\alpha^{-1}(\mathbf{v}) = \mathbf{u}_\alpha^{-1}\left(\alpha \mathbf{H}^\top \mathbf{v} + \mathbf{1}_D \right) \in \mathbb{S}^d, \tag{4}$$

where

$$\mathbf{u}_\alpha^{-1}(\mathbf{x}) = \left(\frac{x_1^{1/\alpha}}{\sum_{j=1}^D x_j^{1/\alpha}}, \ldots, \frac{x_D^{1/\alpha}}{\sum_{j=1}^D x_j^{1/\alpha}} \right). \tag{5}$$

If one is willing to exclude from the sample space the boundary of the simplex, which corresponds to observations that have one or more components equal to zero, then the α-transformation (1) and its inverse (4) are well defined for all $\alpha \in \mathbb{R}$. (Excluding the boundary is standard practise in LRA because the definition is used to sidestep the problem of having data with zeros.) The motivation for transformation (1) is that the case $\alpha = 0$ corresponds to LRA, whereas $\alpha = 1$ corresponds to EDA. We define the case $\alpha = 0$ in terms of the limit $\alpha \to 0$; then

$$\mathbf{z}_0(\mathbf{x}) = \lim_{\alpha \to 0} \mathbf{z}_\alpha(\mathbf{x}) = \mathbf{H} \cdot \mathbf{w}(\mathbf{x}), \tag{6}$$

where

$$\mathbf{w}(\mathbf{x}) = \left(\log \left\{ \frac{x_1}{g(\mathbf{x})} \right\}, \ldots, \log \left\{ \frac{x_D}{g(\mathbf{x})} \right\} \right)^T, \tag{7}$$

is Aitchison's centred logratio transformation (Aitchison, 1983, 2003) and $g(\mathbf{x}) = \prod_{i=1}^D x_i^{1/D}$ is the geometric mean of the components of $\mathbf{x}$. See the Appendix for proof of (6). For the case $\alpha = 1$, (1) is just a linear transformation of the simplex.

Power transformations similar to (1) were considered by Greenacre (2009) and Greenacre (2011), in the somewhat different context of correspondence analysis. A Box–Cox transformation applied to each component of $\mathbf{x} \in \mathbb{S}^d$ so that $\mathbf{x}$ is transformed to

$$\left(\theta^{-1} \left(x_1^\theta - 1 \right), \ldots, \theta^{-1} \left(x_D^\theta - 1 \right) \right)^T, \tag{8}$$

has the limit $(\log x_1, \ldots, \log x_D)^T$ as $\theta \to 0$. We favour transformation (1) in this work in view of its closer connection, via (6), to Aitchison's centred logratio transformation.

The α-transformation (1) leads to a natural simplicial distance measure $\Delta_\alpha(\mathbf{x}, \mathbf{y})$, which we call the α-metric, between observations $\mathbf{x}, \mathbf{y} \in \mathbb{S}^d$, defined in terms of the Euclidean distance $\|\cdot\|$ between transformed observations, i.e.,

$$\begin{aligned} \Delta_\alpha(\mathbf{x}, \mathbf{y}) &= \|\mathbf{z}_\alpha(\mathbf{x}) - \mathbf{z}_\alpha(\mathbf{y})\| \\ &= \frac{D}{|\alpha|} \left[\sum_{i=1}^D \left(\frac{x_i^\alpha}{\sum_{j=1}^D x_j^\alpha} - \frac{y_i^\alpha}{\sum_{j=1}^D y_j^\alpha} \right)^2 \right]^{1/2}. \end{aligned} \tag{9}$$

The special case

$$\Delta_0(\mathbf{x}, \mathbf{y}) := \lim_{\alpha \to 0} \Delta_\alpha(\mathbf{x}, \mathbf{y}) = \left[\sum_{i=1}^D \left(\log \frac{x_i}{g(\mathbf{x})} - \log \frac{y_i}{g(\mathbf{y})} \right)^2 \right]^{1/2}, \tag{10}$$

is Aitchison's distance measure (Aitchison et al., 2000), whereas

$$\Delta_1(\mathbf{x},\mathbf{y}) = D\left[\sum_{i=1}^{D}(x_i - y_i)^2\right]^{1/2} \tag{11}$$

is just Euclidean distance multiplied by D.

Transformation (1), and the implied distance measure (9), offer flexibility in data analysis: the choice of α enables either LRA or EDA, or a compromise between the two, and the particular value of α can be chosen to optimise some measure of practical performance (in this paper, the out-of-sample classification error rate). Crucially, for $\alpha > 0$, the transformation and distance are well defined even when some components have zero values, in contrast to (7) and (10).

Amongst the criteria for compositional distance measures listed by Aitchison (1992), the distance measure (9) satisfies "positivity" ($\Delta_\alpha(\mathbf{x},\mathbf{y}) > 0$ for $\mathbf{x} \neq \mathbf{y}$), "zero difference for equivalent compositions" ($\Delta_\alpha(\mathbf{x},\mathbf{x}) = 0$), "interchangeability of compositions" ($\Delta_\alpha(\mathbf{x},\mathbf{y}) = \Delta_\alpha(\mathbf{y},\mathbf{x})$), "scale invariance" ($\Delta_\alpha(a\mathbf{x}, A\mathbf{y}) = \Delta_\alpha(\mathbf{x},\mathbf{y})$ for all $a > 0$, $A > 0$) and "permutation invariance" ($\Delta_\alpha(P\mathbf{x}, P\mathbf{y}) = \Delta_\alpha(\mathbf{x},\mathbf{y})$ for any permutation P). It does not satisfy "perturbation invariance", a property strongly tied to the logratio transformation (Aitchison, 2003); and nor does it satisfy "subcompositional coherence", a criterion that affects inferences regarding the relationships between compositional components (Greenacre, 2011). The question of how much importance should be given to subcompositional coherence in compositional data analysis has been a matter of much debate; see for example the historical review and discussion in Scealy and Welsh (2014). Our view is similar to that of Scealy and Welsh (2014), which is that subcompositional dominance is not a property of primary importance, although we point out that a referee strongly disagrees with our position. We reiterate that our motivation is to achieve strong practical performance, whether or not our distance measure satisfies any particular properties.

3 Classification techniques for compositional data

The key idea now is to use the α-transformation (1) in conjunction with regularised descriminant analysis (RDA), and the α-metric (9) in conjunction with k-nearest-neighbours (k-NN) classification, to investigate how performance for various values of α compares with the special cases of EDA ($\alpha = 1$) and LRA ($\alpha = 0$). We will begin with a brief review of regularised discriminant analysis, of which linear and quadratic discriminant analysis are special cases, and with the k-nearest neighbours algorithm.

3.1 Regularised discriminant analysis (RDA)

In discriminant analysis we allocate an observation to the group with the highest (posterior) density, assuming that observations in each group come from a multivariate normal distribution. Given a training sample with g groups containing $n_1, \ldots, n_g$ observations, then a new observation $\mathbf{z} \in \mathbb{R}^d$ is classified to the group for which the discriminant score, $\delta_i(\mathbf{z})$, is largest, where

$$\delta_i(\mathbf{z}) = -\frac{1}{2}\log\left|2\pi\hat{\boldsymbol{\Sigma}}_i\right| - \frac{1}{2}(\mathbf{z}-\hat{\boldsymbol{\mu}}_i)^T\hat{\boldsymbol{\Sigma}}_i^{-1}(\mathbf{z}-\hat{\boldsymbol{\mu}}_i) + \log\pi_i; \tag{12}$$

here $|\cdot|$ denotes determinant, $\pi_i = n_i/n$ with $n = \sum_{i=1}^{g} n_i$, and the $\hat{\boldsymbol{\mu}}_i$ and $\hat{\boldsymbol{\Sigma}}_i$ are the sample mean vector and covariance matrix, respectively, of the ith group. Equation (12) is the Bayesian version of discriminant analysis, incorporating the prior group membership probabilities $\boldsymbol{\pi} = (\pi_1, \ldots, \pi_g)$, which assumes that the proportions of observations in the training sample are representative of the proportions in the population. Other choices of $\boldsymbol{\pi}$ are possible depending on available prior information. The frequentist version uses instead $\pi_i = 1/g$. We will use the Bayesian version with $\pi_i = n_i/n$ in our numerical investigations in §4.

The boundary between classification regions, say between groups i and j, is defined by $\delta_i(\mathbf{z}) = \delta_j(\mathbf{z})$. From (12), the boundaries are hence quadratic in $\mathbf{z}$, and for this reason the approach is termed quadratic discriminant analysis (QDA). If we make the simplifying assumption that the groups share a common covariance matrix, then the $\hat{\boldsymbol{\Sigma}}_i$ in (12) can be replaced with the pooled estimate

$$\hat{\boldsymbol{\Sigma}}_p = \frac{\sum_{i=1}^{g} (n_i - 1)\, \hat{\boldsymbol{\Sigma}}_i}{n - g}.$$

In this case, the boundaries are linear, and the approach is hence termed linear discriminant analysis (LDA).

QDA and LDA are special cases of so-called regularised discriminant analysis (RDA); see Hastie et al. (2001, pp. 112-113). The idea of RDA is to regularise the covariance matrices by replacing them with weighted averages

$$\begin{aligned} \hat{\boldsymbol{\Sigma}}_i(\lambda, \gamma) &= \lambda \hat{\boldsymbol{\Sigma}}_i + (1 - \lambda)\, \hat{\boldsymbol{\Sigma}}_p(\gamma), \\ \text{and} \quad \hat{\boldsymbol{\Sigma}}_p(\gamma) &= \gamma \hat{\boldsymbol{\Sigma}}_p + (1 - \gamma)\, \mathrm{tr}\left(\hat{\boldsymbol{\Sigma}}_p\right) \mathbf{I}/d, \end{aligned} \tag{13}$$

where $\lambda, \gamma \in [0, 1]$ are two free parameters and $\mathbf{I}$ is the d-by-d identity matrix. Parameter λ offers a trade-off between the more flexibile classification boundaries of QDA and the greater stability of LDA to one or more of the $\hat{\boldsymbol{\Sigma}}_i$ being ill-conditioned. Parameter γ offers further stability if the pooled estimate $\hat{\boldsymbol{\Sigma}}_p$ is itself ill-conditioned. Choosing $\lambda = 1$ gives QDA, whereas choosing $\lambda = 0$ and $\gamma = 1$ gives LDA.

We propose to use RDA with data transformed using the α-transformation (1), and will denote this by RDA$(\alpha, \lambda, \gamma)$. Hence, RDA$(0, \lambda, \gamma)$ amounts to the LRA approach of applying RDA to data transformed using the isometric log-ratio transformation (6), whereas RDA$(1, \lambda, \gamma)$ amounts to the EDA approach of applying RDA to untransformed data. We will also use the notation QDA(α) = RDA$(\alpha, 1, 0)$ and LDA(α) = RDA$(\alpha, 0, 1)$.

3.2 k-nearest neighbours (k-NN)

The k-NN algorithm is an intuitive classifier that assumes no parametric model. It involves determining the k observations in the training sample that are closest, by some choice of distance measures, to the new test observation, then allocating the test observation to the group most common amongst these k "nearest neighbours". Ties caused by two or more groups jointly being most common can be broken by allocating uniformly at random amongst the tied groups (the strategy we use in our numerical examples in §4) or else by using a secondary tie-breaking criterion.

Performance of k-NN depends of the choice of k: small k allows for classification boundaries

which are flexible but which have a tendency to overfit, with the opposites true when k is large. It also depends on the choice of distance measure. Since we are dealing with compositional data we shall use the α-metric (9), denoting such an approach k-NN(α), so k-NN(0) indicates the LRA approach of using k-NN with Aitchison's distance (10), while k-NN(1) indicates the EDA approach of using k-NN based on Euclidean distance.

We can equally easily use any of many other possible distance measures. For sake of comparing performance with the α-metric we also consider one alternative, namely the following variant of the Jensen-Shannon divergence:

$$\mathrm{ESOV}(\mathbf{x},\mathbf{y}) = \sqrt{\sum_{i=1}^{D}\left(x_i \log\frac{2x_i}{x_i+y_i} + y_i \log\frac{2y_i}{x_i+y_i}\right)}. \tag{14}$$

We use the notation ESOV after Endres and Schindelin (2003) and Österreicher and Vajda (2003) who independently proved that (14) satisfies the triangle inequality and thus is a metric. As with the α-metric (9), the ESOV metric (14) is well defined even when zero values are present. We denote the k-NN classifier based on metric (14) by k-NN$_{\mathrm{ESOV}}$.

4 Applications of compositional classification

We will show four examples of applications of the proposed compositional discrimination techniques. In all cases we used real data sets, two of them having observations with zero values in some of the components, and the other two data sets having no zero values. The two benchmarks for comparison will be when $\alpha = 0$, which results in LRA, and when $\alpha = 1$, which results in EDA.

We performed RDA(α, λ, γ), k-NN(α), varying α in steps of 0.05 between -1 and 1 for datasets without zeros and between 0.05 and 1 for datasets with zeros (since in such circumstances the α-transformation and α-metric are not defined for $\alpha \leq 0$), and varying the values of λ and γ in steps of 0.1 between 0 and 1.

To estimate the rate of correct classification in out-of-sample prediction we used cross validation. This involves dividing the set of n observations into training and test sets of size n_{train} and n_{test} respectively, training the classifier on the training set, then evaluating its prediction accuracy of the test set. In view of the samples having groups with quite variable numbers of observations we used stratified random sampling to ensure that the training sets were representative of the test sets, and to arrange that all groups were represented in the test set. In particular, we randomly divided the samples into training and test sets so that

$$\frac{n_i}{n} \approx \frac{n_{i,\mathrm{train}}}{n_{\mathrm{train}}} \approx \frac{n_{i,\mathrm{test}}}{n_{\mathrm{test}}},$$

where n_i, $n_{i,\mathrm{train}}$ and $n_{i,\mathrm{test}}$ are the sample sizes of the ith group in the full, training and test samples, respectively. We then estimated the rate of correct classification by

$$q = \frac{c}{n_{\mathrm{test}}}, \tag{15}$$

where c is the number of observations in the test set correctly classified and n_{test} is the test sample size.

For each of the classifiers RDA(α, λ, γ) and k-NN(α), the steps can be summarised as follows

Step 1. Partition the sample into training and test sets using stratified random sampling.

Step 2. For each combination of values of the free parameters (α, λ, γ for RDA; α, k for k-NN(α); train the classifier on the training set.

Step 3. Apply the classifiers to the test set, and calculate q in (15).

Step 4. Repeat steps $1-3$ a large number, say B, times, then estimate the rate of correct classification as the average of the qs in Step 3.

For the calculations in the following section we took $B = 200$ which gave estimates of the rate of correct classification with small standard errors at reasonable computational cost.

4.1 Examples

We will now introduce four datasets to investigate the performance of the supervised classification techniques described in §3. The datasets come from different fields, namely ecology, forensic science, hydrochemistry and economics.

Example 1: Fatty acid signature data (contains zero values)

This is a dataset described in (Stewart and Field, 2011) (itself an updated version of a dataset from (Iverson et al., 2004)) which contains observations of $n = 2110$ fish of $g = 28$ different species, each observation being a composition with $D = 40$ components that characterises the fatty acid signature of the fish. A special feature of this dataset is that it contains many zero values (3506 components, across all observations, are zero) which rules out use of the log-ratio transformation (7). Table 1 shows the number of observations in each group, and the number of observations for which at least one component is zero. Table 2 shows the proportion of observations which have zeros in each of the components. For this example, for the cross validation we used a test set of $n_{\text{test}} = 165$ observations (7.8% of the full sample).

Species	Sample size	Species	Sample size	Species	Sample size
Butterfish	35(30)	Mackerel	34(23)	Snake Blenny	18(12)
Capelin	165(145)	Ocean Pout	31(31)	Squid	18(17)
Cod	147(121)	Plaice	148(120)	Thorny Skate	74(74)
Gaspereau	70(69)	Pollock	57(49)	Turbot	20(20)
Haddock	148(134)	Red Hake	25(24)	White Hake	75(71)
Halibut	13(11)	Redfish	84(74)	White Flounder	90(80)
Herring	247(231)	Sandlance	124(101)	Winter Skate	40(39)
Lobster	21(21)	Shrimp	122(87)	Witch Flounder	24(24)
Longhorn Sculpin	70(69)	Silver Hake	70(58)	Yellow Tail	118(103)
Lumpfish	22(13)				

Table 1: Sample sizes of the different species of the fatty acid data. The number inside the parentheses shows how many observations have at least one zero element.

Component	1st	2nd	3rd	4th	5th	6th	7th	8th	9th	10th
Percentage of zeros	0.00%	0.00%	0.00%	6.54%	0.28%	9.86%	9.10%	4.88%	65.36%	2.94%
Component	11th	12th	13th	14th	15th	16th	17th	18th	19th	20th
Percentage of zeros	0.00%	0.00%	0.00%	0.00%	6.78%	2.32%	0.62%	0.00%	3.51%	0.05%
Component	21st	22nd	23rd	24th	25th	26th	27th	28th	29th	30th
Percentage of zeros	2.65%	0.09%	0.00%	0.05%	1.80%	1.66%	0.00%	0.33%	0.05%	0.00%
Component	31st	32nd	33rd	34th	35th	36th	37th	38th	39th	40th
Percentage of zeros	0.33%	0.5%	0.00%	27.35%	0.00%	10.66%	0.00%	8.91%	0.00%	0.00%

Table 2: Fatty acid data: the percentage of observations for which each component is zero.

Example 2: Forensic glass data (contains zero values)

In the second example we use the forensic glass dataset (UC Irvine Machine Learning Repository, 2014) which has $n = 214$ observations from $g = 6$ different categories of glass, where each observation is a composition with $D = 8$ components. The categories which occur are: containers (13 observations, 12 of which have at least one zero element), vehicle headlamps (29 observations, all with at least one zero value), tableware (9 observations, all with at least one zero value), vehicle window glass (17 observations, 16 with at least one zero value), window float glass (70 observations, 69 with at least one zero value) and window non-float glass (76 observations, 72 with at least one zero value). Once again the zeros rule out the use of LRA. In total there are 392 zero values; Table 3 shows in which components these zeros arise and Table 6 summarises the distribution of zeros across the observations. For the cross validation we used a test set consisted of $n_{\text{test}} = 30$ observations (14% of the total sample).

Components	Sodium	Magnesium	Aluminium	Silicon
Percentage of zeros	0.00%	19.63%	0.00%	0.00%
Components	Potassium	Calcium	Barium	Iron
Percentage of zeros	14.02%	0.00%	82.24%	67.29%

Table 3: Forensic glass data: the percentage of observations for which each component is zero.

Example 3: Hydrochemical data (contains no zero values)

The hydrochemical data set (Otero et al., 2005) contains compositional observations on $D = 14$ chemicals (H, Na, K, Ca, Mg, Sr, Ba, NH4, Cl, HCO3, NO3, SO4, PO4, TOC) in water samples from tributaries of the Llobregat river in north-east Spain. The $n = 485$ observations are in $g = 4$ groups according to which tributary they were measured in: Anoia (143 observations), Cardener (95 observations), Upper Llobregat (135 observations) or Lower Llobregat (112 observations). For the cross validation in this example we used a training set of size $n_{\text{test}} = 165$ (34% of the total sample size).

Example 4: National income data (contains no zero values)

This final example is an economics data set (Larrosa, 2003) containing compositional observations for $n = 56$ countries with $D = 5$ components reflecting the proportion of capital allocated

in production assets, residential buildings, non-residential buildings, other buildings, and transportation equipment. The countries are categorised into $g = 5$ groups according to income levels and membership of the Organization for Economic Co-operation and Development (OECD); the groups are "low income" (10 countries), "lower middle income" (12 countries), "upper middle income" (9 countries), "high income and OECD member" (21 countries), and "high income and non-OECD member" (4 countries). For the cross validation, we used a test set of $n_{\text{test}} = 10$ observations (17.9% of the total sample).

4.2 Results

This section contains results from applying the methods of §3 to the four compositional datasets described above. Results are summarised in Figures 1 and 2 and Tables 4-7. The Tables show results for $\alpha = 1$, $\alpha = 0$, and for α free in [-1,1], in each case for the values of free parameters that maximise the estimated rate of correct classification.

Example 1 (Fatty acid signature data)

Method	Estimated rate of correct classification	Method	Estimated rate of correct classification
RDA$(0.6, 0.9, 0.7)$	0.962(0.014)	RDA$(1, 0.8, 1)$	0.949(0.016)
LDA(0.45)	0.897(0.022)	LDA(1)	0.868(0.024)
2-NN(0.35)	0.933(0.020)	2-NN(1)	0.849(0.027)
2-NN$_{ESOV}$	0.921(0.019)		

Example 2 (Forensic glass data)

Method	Estimated rate of correct classification	Method	Estimated rate of correct classification
RDA$(0.95, 0.1, 1)$	0.643(0.034)	RDA$(1, 0.1, 1)$	0.643(0.034)
LDA(0.4)	0.629(0.034)	LDA(1)	0.629(0.034)
3-NN(0.85)	0.719(0.033)	2-NN(1)	0.719(0.033)
3-NN$_{ESOV}$	0.693(0.033)		

Table 4: Estimated rate of correct classification of the different approaches. The standard error appears inside the parentheses.

	Number of zeros				
Method	0 (12.27%)	1 (45.69%)	2 (2.61%)	3 (9.38%)	4-8 (10.5%)
RDA$(0.6, 0.90.71)$	0.956(0.043)	0.963(0.021)	0.963(0.030)	0.976(0.040)	0.949(0.053)
RDA$(1, 0.8, 1)$	0.951(0.045)	0.962(0.022)	0.941(0.037)	0.942(0.063)	0.911(0.066)
LDA(0.45)	0.875(0.065)	0.925(0.029)	0.881(0.051)	0.872(0.091)	0.855(0.093)
LDA(1)	0.882(0.060)	0.898(0.034)	0.842(0.053)	0.822(0.099)	0.813(0.096)
2-NN(0.35)	0.923(0.054)	0.938(0.025)	0.923(0.039)	0.963(0.047)	0.922(0.064)
2-NN(1)	0.844(0.075)	0.853(0.036)	0.853(0.058)	0.874(0.082)	0.803(0.100)
2-NN$_{ESOV}$	0.918(0.062)	0.928(0.030)	0.913(0.046)	0.962(0.050)	0.880(0.086)

Table 5: Fatty acid signature data: classification accuracy by number of zeros. The estimated rate of correct classification is shown (with standard errors in parentheses).

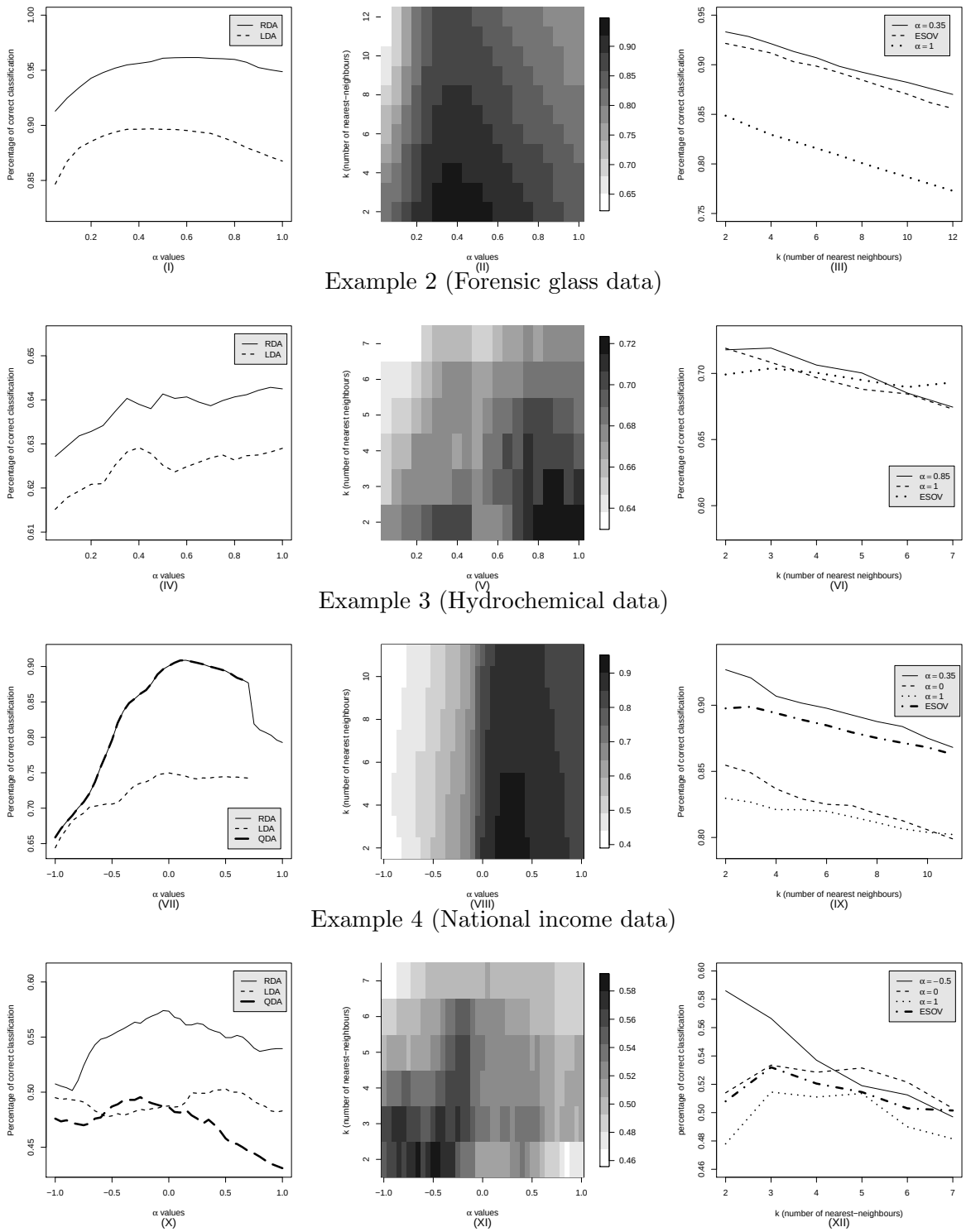
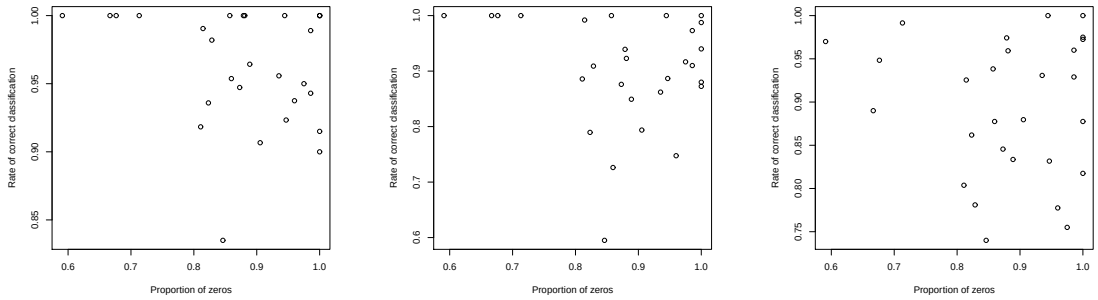


Figure 1: All graphs contain the estimated rate of correct classification for the different methods. The first column refers to LDA, QDA and RDA as a function of α. The second column contains the heat plots of the k-NN algorithm as a function of α and k, the nearest neighbours. The graphs in the third column present the results of the k-NN algorithm of the α-metric for some specific values of α.



RDA$(0.6, 0.9, 0.7)$ LDA(0.45) 2-NN(0.35)

Figure 2: Fatty acid signature data: the estimated rate of correct classification accuracy by group versus the the proportion of observations within the group that contain at least one zero.

	Number of zeros				
Method	0 (3.27%)	1 (29.44%)	2 (50.93%)	3 (13.55%)	4 (2.80%)
RDA$(0.95, 0.1, 1)$	0.421(0.433)	0.582(0.173)	0.668(0.111)	0.787(0.233)	0.402(0.463)
RDA$(1, 0.1, 1)$	0.428(0.435)	0.585(0.174)	0.665(0.112)	0.788(0.233)	0.397(0.459)
LDA(0.4)	0.404(0.431)	0.536(0.165)	0.636(0.120)	0.869(0.194)	0.689(0.412)
LDA(1)	0.397(0.430)	0.523(0.162)	0.673(0.110)	0.790(0.230)	0.463(0.462)
3-NN(0.85)	0.307(0.394)	0.713(0.160)	0.717(0.108)	0.925(0.146)	0.387(0.429)
2-NN(1)	0.568(0.441)	0.715(0.160)	0.712(0.114)	0.870(0.178)	0.387(0.429)
3-NN$_{ESOV}$	0.477(0.447)	0.644(0.164)	0.764(0.097)	0.731(0.243)	0.387(0.429)

Table 6: Forensic glass data: classification accuracy by number of zeros. The estimated rate of correct classification is shown (with standard errors in parentheses).

Fatty acid and glass data from Examples 1 and 2

For both the fatty acid and forensic glass datasets, some of the groups have fewer observations than the dimension D of the compositions, so QDA cannot be applied (since at least one of the $\hat{\mathbf{\Sigma}}_i$ in (12) is singular). Both RDA, LDA and k-NN are applicable, however, and Table 4 shows a comparison of performance for these techniques.

For the fatty acid data, RDA performs strongest, and best performance is achieved when $\alpha = 0.6$. For this dataset k-NN(α) performs strongly too, with $\alpha = 0.35$ giving notably better performance than $\alpha = 1$ (which corresponds to the EDA approach). For the forensic glass data, k-NN outperformed RDA and LDA, and the flexibility of having α different from 1 offered no improvement.

For both of these datasets, results suggest that there is no clear relationship between classification accuracy for the groups and the number of observations containing zeros, i.e., no clear evidence that observations with zeros were more or less difficult to classify correctly than those without zeros. Figure 2, for example, shows the classification accuracy for each group in the fatty acid dataset plotted against the proportion of observations that contain at least one zero, and no clear correlation is apparent. Results (not shown) for the k-NN with the ESOV metric (14) similarly show little pattern. Table 5 shows results for the fatty acid data presented according to the number of zeros in the observations. There is no clear relationsip between

Example 3 (Hydrochemical data)

Method	Estimated rate of correct classification	Method	Estimated rate of correct classification	Method	Estimated rate of correct classification
RDA$(0.15, 1, 0)$	0.909(0.02)	RDA$(0, 1, 0)$	0.901(0.021)	RDA$(1, 0.9, 0.9)$	0.793(0.029)
QDA(0.15)	0.909(0.02)	QDA(0)	0.901(0.021)	QDA(1)	-
LDA(0)	0.750(0.031)	LDA(0)	0.750(0.031)	LDA(1)	-
2-NN(0.25)	0.927(0.020)	2-NN(0)	0.855(0.026)	2-NN(1)	0.830(0.027)
3-NN$_{ESOV}$	0.899(0.021)				

Example 4 (National income data)

Method	Estimated rate of correct classification	Method	Estimated rate of correct classification	Method	Estimated rate of correct classification
RDA$(-0.05, 0.5, 0)$	0.574(0.035)	RDA$(0, 0.5, 0)$	0.574(0.035)	RDA$(1, 0.2, 0)$	0.540(0.035)
QDA(-0.25)	0.496(0.035)	QDA(0)	0.487(0.035)	QDA(1)	0.431(0.035)
LDA(0.5)	0.503(0.035)	LDA(0)	0.488(0.035)	LDA(1)	0.483(0.035)
2-NN(-0.5)	0.586(0.035)	3-NN(0)	0.533(0.035)	3-NN(1)	0.515(0.035)
3-NN$_{ESOV}$	0.541(0.035)				

Table 7: Estimated rate of correct classification of the different approaches (with standard errors in parentheses).

classification accuracy and number of zeros. Corresponding results in Table 6 for the forensic glass data show lower classification accuracy for observations with 0 or 4 zeros compared with observations with 1, 2 or 3 zeros, albeit with large standard errors on account of the small number of such observations. Hence, again, the conclusion is that there is no clear evidence that zeros make observations any more or less difficult to classify correctly.

The key points from these examples are that LRA is not directly applicable because of the zeros, but EDA ($\alpha = 1$) performs quite well with RDA having better performance in one example and k-NN in another, and in one of the examples letting α be a value other than 1 gave a further improvement.

Hydrochemical and national income data from Examples 3 and 4

For the hydrochemical data the extra flexibility of RDA over QDA offers no improvement (and hence RDA and QDA give identical results). Ill-conditioning of covariance matrices makes QDA and LDA unstable for $\alpha > 0.75$, which is why in Figure 1(VII) the lines corresponding to these methods stop at $\alpha = 0.75$. The plots in the left column of Figure 1 show clearly that the performance of RDA$(\alpha, \lambda, \gamma)$ (and its special cases QDA(α) and LDA(α)) depends on α and tend to do best at values of α other than 0 or 1. k-NN(α) does best for this example, with $\alpha = 0.25$ and 2 nearest neighbours, leading to the best performance of all the classifiers.

For the final example of the national income data, the LRA approach of taking $\alpha = 0$ leads to the best performance of RDA. As in the previous example the k-NN classifier does best when $\alpha = -0.5$ and 2 neighbours are considered.

5 Conclusions

We have considered the α-transformation (1) and the α-metric (9) as a means to adapt LDA, QDA, RDA and k-NN for compositional data. This generalises EDA and LRA approaches via the parameter α, the choice of which enable a compromise between the two. Rather than choosing either EDA or LRA, our approach enables a choice of α based on the dataset at hand, and numerical results suggest there is a clear benefit to having this flexibility.

An important benefit is that such an approach is well defined even when the dataset contains observations with components equal to zero, unlike with LRA in which ad hoc modifications to the data are needed prior to applying the log-ratio transformation. Within k-NN it is simple to incorporate any choice of distance that seems appropriate.

Appendix

Relationship between the α-transformation and centred log-ratio transformation

The proof that the transformation $(D\mathbf{u}_\alpha(\mathbf{x}) - \mathbf{1}_D)/\alpha$ defined on the right-hand side of (1) tends to the centred log-ratio transformation (7) as $\alpha \to 0$ is as follows: for component i,

$$\begin{aligned}
\frac{1}{\alpha}\left(\frac{Dx_i^\alpha}{\sum_{j=1}^D x_j^\alpha} - 1\right) &= \frac{D}{\alpha}\left[\frac{1 + \alpha \log x_i + O\left(\alpha^2\right)}{D\left(1 + \frac{\alpha}{D}\sum_{j=1}^D \log x_j + O\left(\alpha^2\right)\right)} - \frac{1}{D}\right] \\
&= \left\{(1 + \alpha \log x_i)\left(1 + \frac{\alpha}{D}\sum_{j=1}^D \log x_j\right)^{-1} - 1 + O\left(\alpha^2\right)\right\} \Big/ \alpha \\
&= \left\{(1 + \alpha \log x_i)\left(1 - \frac{\alpha}{D}\sum_{j=1}^D \log x_j\right) - 1 + O\left(\alpha^2\right)\right\} \Big/ \alpha \\
&= \left\{1 + \alpha \log x_i - \frac{\alpha}{D}\sum_{j=1}^D \log x_j - 1 + O\left(\alpha^2\right)\right\} \Big/ \alpha \\
&= \log x_i - \log \prod_{j=1}^D x_j^{1/D} + O\left(\alpha\right) \\
&\to \log\left\{\frac{x_i}{g\left(\mathbf{x}\right)}\right\} \quad \text{as} \quad \alpha \to 0.
\end{aligned}$$

The proof that the α-metric (9) tends to the LRA metric (10) as $\alpha \to 0$ follows from this proof.

References

AITCHISON, J. (1982), "The statistical analysis of compositional data", *Journal of the Royal Statistical Society. Series B, 44,* 139–177.

AITCHISON, J. (1983), "Principal component analysis of compositional data", *Biometrika, 70,* 57–65.

AITCHISON, J. (1992), "On criteria for measures of compositional difference", *Mathematical Geology, 24,* 365–379.

AITCHISON, J. (2003), "The Statistical Analysis of Compositional Data" (Reprinted with additional material by The Blackburn Press), London (UK): Chapman & Hall.

AITCHISON, J. and BARCELO-VIDAL, C. and MARTIN-FERNANDEZ, J.A. and PAWLOWSKY-GLAHN, V. (2000), "Logratio analysis and compositional distance", *Mathematical Geology, 32,* 271–275.

BAXTER, M. J. (2001), "Statistical modelling of artefact compositional data", *Archaeometry, 43,* 131–147.

BAXTER, M. J., BEARDAH, C. C., COOL, H. E. M., and JACKSON, C. M. (2005), "Compositional data analysis of some alkaline glasses", *Mathematical Geology, 37,* 183–196.

BAXTER, M. J. and FREESTONE, I. C. (2006), "Log-ratio compositional data analysis in archaeometry", *Archaeometry, 48,* 511–531.

BUTLER, A. and GLASBEY, C. (2008), "A latent Gaussian model for compositional data with zeros", *Journal of the Royal Statistical Society: Series C, 57,* 505–520.

DRDYEN, I. L. and MARDIA, K. V. (1998), "Statistical Shape Analysis", New York: Wiley.

EGOZQUE, J.J. and PAWLOWSKY-GLAHN, V. and MATEU-FIGUERAS, G. and BARCELO-VIDAL, C. (2003), "Isometric logratio transformations for compositional data analysis", *Mathematical Geology, 35,* 279–300.

ENDRES, D. M. and SCHINDELIN, J. E. (2003), "A new metric for probability distributions", *Information Theory, IEEE Transactions on, 49,* 1858–1860.

FRY, J. M., FRY, T. R. L., and McLAREN, K. R. (2000), "Compositional data analysis and zeros in micro data", *Applied Economics, 32,* 953–959.

GREENACRE, M. (2009), "Power transformations in correspondence analysis", *Computational Statistics & Data Analysis, 53,* 3107–3116.

GREENACRE, M. (2011), "Measuring subcompositional incoherence", *Mathematical Geosciences, 43,* 681–693.

GUEORGUIEVA, R., ROSENHECK, R., and ZELTERMAN, D. (2008), "Dirichlet component regression and its applications to psychiatric data", *Computational statistics & data analysis, 52,* 5344–5355.

HASTIE, T., TIBSHIRANI, R., and FRIEDMAN, J. (2001), "The Elements of Statistical Learning: Data Mining, Inference, and Prediction", Berlin: Springer.

IVERSON, S. J., FIELD, C., Don BOWEN, W., and BLANCHARD, W. (2004), "Quantitative fatty acid signature analysis: a new method of estimating predator diets", *Ecological Monographs, 74,* 211–235.

LANCASTER, H. O. (1965), "The Helmert matrices", *American Mathematical Monthly, 72,* 4–12.

LARROSA, J. M. (2003), "A compositional statistical analysis of capital stock", *In Proceedings of the 1st Compositional Data Analysis Workshop,* Girona, Spain.

NEOCLEOUS, T., AITKEN, C., and ZADORA, G. (2011), "Transformations for compositional data with zeros with an application to forensic evidence evaluation", *Chemometrics and Intelligent Laboratory Systems, 109,* 77–85.

OSTERREICHER, F. and VAJDA, I. (2003), "A new class of metric divergences on probability spaces and its applicability in statistics", *Annals of the Institute of Statistical Mathematics, 55,* 639–653.

OTERO, N., TOLOSANA-DELGADO, R., SOLER, A., PAWLOWSKY-GLAHN, V., and CANALS, A. (2005), "Relative vs. absolute statistical analysis of compositions: A comparative study of surface waters of a mediterranean river", *Water Research, 39,* 1404–1414.

PALAREA-ALBALADEJO, J., MARTIN-FERNANDEZ, J. A. and SOTO, J. A. (2012), "Dealing with distances and transformations for fuzzy c-means clustering of compositional data", *Journal of classification, 29,* 144–169.

RODRIGUES, P. C. and LIMA, A. T. (2009), "Analysis of an European union election using principal component analysis", *Statistical Papers, 50,* 895–904.

SCEALY, J. L. and WELSH, A. H. (2011), "Regression for compositional data by using distributions defined on the hypersphere", *Journal of the Royal Statistical Society. Series B, 73,* 351–375.

SCEALY, J. L. and WELSH, A. H. (2014), "Colours and cocktails: compositional data analysis. 2013 Lancaster lecture", *Australian & New Zealand Journal of Statistics, 56,* 145–169.

STEPHENS, M. A. (1982), "Use of the von Mises distribution to analyse continuous proportions", *Biometrika, 69,* 197–203.

STEWART, C. and FIELD, C. (2011), "Managing the essential zeros in quantitative fatty acid signature analysis", *Journal of Agricultural, Biological, and Environmental Statistics, 16,* 45–69.

TSAGRIS, M. T., PRESTON, S., and WOOD, A. T. A. (2011), "A data-based power transformation for compositional data", *In Proceedings of the 4th Compositional Data Analysis Workshop,* Girona, Spain.

UC IRVINE MACHINE LEARNING REPOSITORY (2014), "Forensic Glass Dataset", *http://archive.ics.uci.edu/ml/datasets/Glass+Identification.*

WORONOW, A. (1997), "The elusive benefits of logratios", *In Proceedings of the 3rd Annual Conference of the International Association for Mathematical Geology,* Barcelona, Spain.

ZADORA, G., NEOCLEOUS, T., and AITKEN, C. (2010), "A two-level model for evidence evaluation in the presence of zeros", *Journal of forensic sciences, 55,* 371–384.