Flexible non-parametric regression models for compositional response data with zeros

Michail Tsagris, Abdulaziz Alenazi and Connie Stewart

Department of Economics, University of Crete, Greece
mtsagris@uoc.gr
Department of Mathematics, College of Science, Northern Border University, Saudi Arabia
a.alenazi@nbu.edu.sa
Department of Mathematics and Statistics, University of New Brunswick, Saint John, Canada
Connie.Stewart@unb.ca

February 3, 2023

Abstract

Compositional data arise in many real-life applications and versatile methods for properly analyzing this type of data in the regression context are needed. When parametric assumptions do not hold or are difficult to verify, non-parametric regression models can provide a convenient alternative method for prediction. To this end, we consider an extension to the classical k–NN regression, termed α–k–NN regression, that yields a highly flexible non-parametric regression model for compositional data through the use of the α-transformation. Our model is further extended to the α–kernel regression by adopting the Nadaraya-Watson estimator. Unlike many of the recommended regression models for compositional data, zeros values (which commonly occur in practice) are not problematic and they can be incorporated into the proposed models without modification. Extensive simulation studies and real-life data analyses highlight the advantage of using these non-parametric regressions for complex relationships between the compositional response data and Euclidean predictor variables. Both suggest that α–k–NN and α–kernel regressions can lead to more accurate predictions compared to current regression models which assume a, sometimes restrictive, parametric relationship with the predictor variables. In addition, the α–k–NN regression, in contrast to current regression techniques, enjoys a high computational efficiency rendering it highly attractive for use with large scale, massive, or big data.

Keywords: compositional data, regression, α–transformation, k–NN algorithm, kernel regression

1 Introduction

Non-negative multivariate vectors with variables (typically called components) conveying only relative information are referred to as compositional data. When the vectors are normalized to sum to 1, their sample space is the standard simplex given below

$$\mathbb{S}^{D-1} = \left\{ (u_1, ..., u_D)^\top \middle| u_i \geq 0, \sum_{i=1}^{D} u_i = 1 \right\}, \tag{1}$$

where D denotes the number of components.

Examples of compositional data may be found in many different fields of study and the extensive scientific literature that has been published on the proper analysis of this type of data is indicative of its prevalence in real-life applications[1]. It is perhaps not surprising, given the widespread occurrence of this type of data, that many compositional data analysis applications involve covariates. In sedimentology, for example, samples were collected from an Arctic lake and the change in their chemical composition across different water depths was of interest (van den Boogaart et al., 2018). This data set is analyzed in Section 5 using our proposed methodology along with several other data sets. These include compositional glacial data, household consumption expenditures data, data on the concentration of chemical elements in samples of soil, data on morphometric measurements of fish, as well as electoral, pollution and power data, all of which are associated with some covariates. In addition to these examples, the literature cites numerous other applications of compositional regression analysis. For example, data from oceanography studies involving Foraminiferal (a marine plankton species) compositions at different sea depths from oceanography were analyzed in Aitchison (2003). In hydrochemistry, Otero et al. (2005) used regression analysis to draw conclusions about anthropogenic and geological pollution sources of rivers in Spain. In economics, Morais et al. (2018) linked market shares to some independent variables, while in political sciences the percentage of votes of each candidate were linked to some relevant predictor variables (Katz and King, 1999, Tsagris and Stewart, 2018). Finally, in the field of bioinformatics compositional data techniques have been used for analysing microbiome data (Chen and Li, 2016, Shi et al., 2016, Xia et al., 2013).

The need for valid regression models for compositional data in practice has led to several developments in this area, many of which have been proposed in recent years. The first regression model for compositional response data was developed by Aitchison (2003), commonly referred to as Aitchison's model, and was based on the additive log-ratio transformation defined in Section 2. Dirichlet regression was applied to compositional data in Gueorguieva et al. (2008), Hijazi and Jernigan (2009), Melo et al. (2009). Iyengar and Dey (2002) investigated the generalized Liouville family of distributions that permits distributions with negative or mixed correlation and also contains non-Dirichlet distributions with non-positive correlation. The additive log-ratio transformation was again used by Tolosana-Delgado and von Eynatten (2009) while Egozcue et al. (2012) extended Aitchison's regression model by using an isometric log-ratio transformation but, instead of employing the usual Helmert sub-matrix, Egozcue et al. (2012) chose a different orthogonal matrix that is compositional data dependent.

---

[1]For a substantial number of specific examples of applications involving compositional data see (Tsagris and Stewart, 2020).

A drawback of the aforementioned regression models is their inability to handle zero values directly and, consequently, a few models have recently been proposed to address the zero problem. In particular, Scealy and Welsh (2011) transformed the compositional data onto the unit hyper-sphere and introduced the Kent regression which treats zero values naturally. Spatial compositional data with zeros were modelled in Leininger et al. (2013) from the Bayesian stance. In Mullahy (2015), regression models for economic share data were estimated, with the shares taking zero values with nontrivial probability. Alternative regression models in the field of econometrics and applicable when zero values are present are discussed in Murteira and Ramalho (2016). In Tsagris (2015a), a regression model that minimizes the Jensen-Shannon divergence was proposed while in Tsagris (2015b), α-regression (a generalization of Aitchison's log-ratio regression) was introduced, and both of these approaches are compatible with zeros. An extension to Dirichlet regression allowing for zeros was developed by Tsagris and Stewart (2018) and referred to as zero adjusted Dirichlet regression.

Most of the preceding regression models are parametric models in the sense that they are limited to the assumptions of linear or generalized linear relationships between the dependent and independent variables, even though the relationships in many real-life applications are not restricted to the linear setting nor conform to fixed parametric forms. In the case of unconstrained data, when parametric assumptions are not satisfied or easily verified, non-parametric regression models and algorithms, such as k–NN regression, are often considered. As detailed in Section 3, Kernel regression (Wand and Jones, 1994) is a more sophisticated technique that generalizes k–NN regression by adding different weights to each observation that decay exponentially with distance. A disadvantage of Kernel regression is that it is more complex and computationally expensive than k–NN regression.

The contribution of this paper is an extension of these classical non-parametric approaches for application to compositional data through the utilization of the α–transformation. Specifically, the proposed α–k–NN and α–kernel regressions for compositional data link the predictor variables in a non-parametric, non-linear fashion, thus allowing for more flexibility. The models have the potential to provide a better fit to the data compared to conventional models and yield improved predictions when the relationships between the compositional and the non-compositional variables are complex. Furthermore, in contrast to other non-parametric regressions such as projection pursuit, applicable to log-ratio transformed compositional data (see Friedman and Stuetzle (1981)), the two proposed methods allow for zero values in the data. While our objective is improved prediction for complex relationships involving compositional response data, a disadvantage of non-parametric regression strategies, in general, is that the usual statistical inference of the effect of the independent variables is not straightforward. However, the use of ICE plots (Goldstein et al., 2015) can overcome this obstacle and facilitate visualization of the effect of the independent variables. Finally, a significant advantage of α–k–NN regression in particular is its high computational efficiency compared to all current regression techniques, even when the sample sizes number hundreds of thousands or even millions of observations. This is also true for the case of big data and online data or streaming data that require rapid predictions and our work lays the foundation for novel methods in machine learning since many currently rely on k–NN methodology. Functions to carry out both α–k–NN and α–kernel are provided in the R package *Compositional*.

While in practice either the response or predictor data (or both) may be compositional, we consider only the case in which the response data are compositional. The case where the response data are Euclidean and the predictor variables are compositional was examined but the proposed method did not outperform the α–Principal Components Regression in (Tsagris, 2015b). The paper is structured as follows: Section 2 describes relevant transformations and regression models for compositional data, while in Section 3, the proposed α–k–NN and α–kernel regression models are introduced. We examine the advantages and limitations of these models through simulation studies, implemented in Section 4, and real-life data sets are presented in Section 5. Finally, concluding remarks are provided in Section 6.

2 Compositional data analysis: transformations and regression models

In this section some preliminary definitions and methods in compositional data analysis relevant to the work in this paper are introduced. Specifically, two commonly used log-ratio transformations, as well as a more general α–transformation are defined, followed by some existing regression models for compositional regression.

2.1 Transformations

2.1.1 Additive log-ratio transformation

Aitchison (1982) suggested applying the additive log-ratio (alr) transformation to compositional data prior to using standard multivariate data analysis techniques. Let $\mathbf{u} = (u_1, \ldots, u_D)^\top \in \mathbb{S}^{D-1}$, the alr transformation is given by

$$\mathbf{v} = \{v_i\}_{i=1,\ldots,D-1} = \left\{ \log \frac{u_i}{u_1} \right\}_{i=2,\ldots,D}, \tag{2}$$

where $\mathbf{v} = (v_1, \ldots, v_{D-1}) \in \mathbb{R}^{D-1}$. Note that the common divisor, u_1, need not be the first component and was simply chosen for convenience. The inverse of Equation (2) is given by

$$\mathbf{u} = \{u_i\}_{i=1,\ldots,D} = \left\{ \begin{array}{ll} \frac{1}{1+\sum_{j=2}^{D} e^{v_j}}, & i = 1 \\ \frac{e^{v_{i-1}}}{1+\sum_{j=2}^{D} e^{v_j}}, & i = 2, \ldots, D \end{array} \right\}. \tag{3}$$

2.1.2 Isometric log-ratio transformation

An alternative transformation proposed by Aitchison (1983) is the centred log-ratio (clr) transformation defined as

$$\mathbf{y} = \{y_i\}_{i=1,\ldots,D} = \left\{ \log \frac{u_i}{g(\mathbf{u})} \right\}_{i=1,\ldots,D,} \tag{4}$$

where $g(\mathbf{u}) = \prod_{j=1}^{D} u_j^{1/D}$ is the geometric mean. The inverse of Equation (4) is given by

$$\mathbf{u} = \{u_i\}_{i=1,\ldots,D} = \left\{ \frac{e^{y_i}}{\sum_{j=1}^{D} e^{y_j}} \right\}_{i=1,\ldots,D} = \mathcal{C}\{e^{\mathbf{y}}\}, \tag{5}$$

where $\mathcal{C}\{.\}$ denotes the closure operation, or normalization to the unity sum.

The clr transformation in Equation (4) was proposed in the context of principal component analysis and its drawback is that $\sum_{i=1}^{D} y_i = 0$, so essentially the problem of the unity sum constraint is replaced by the problem of the zero sum constraint. In order to address this issue, Egozcue et al. (2003) proposed multiplying Equation (4) by the $(D-1) \times D$ Helmert sub-matrix $\mathbf{H}$ (Dryden and Mardia, 1998, Lancaster, 1965, Le and Small, 1999), an orthogonal matrix with the first row omitted, which results in what is called the isometric log-ratio (ilr) transformation

$$\mathbf{z}_0 = \mathbf{H}\mathbf{y}, \tag{6}$$

where $\mathbf{z}_0 = (z_{01}, \ldots, z_{0D-1})^\top \in \mathbb{R}^{D-1}$. Note that any orthogonal matrix which preserves distances would also be appropriate in place of $\mathbf{H}$ (Tsagris et al., 2011). The inverse of Equation (6) is

$$\mathbf{u}_0 = \mathcal{C}\left\{e^{\mathbf{H}^\top \mathbf{z}_0}\right\}. \tag{7}$$

2.1.3 α−transformation

The main disadvantage of the above transformations is that they do not allow zero values in any of the components, unless a zero value imputation technique (see Martín-Fernández et al. (2003), for example) is first applied. This strategy, however, can produce regression models with predictive performance worse than regression models that handle zeros naturally (Tsagris, 2015a). When zeros occur in the data, the power transformation introduced by Aitchison (2003), and subsequently modified by Tsagris et al. (2011), may be used. Specifically, Aitchison (2003) defined the power transformation as

$$\mathbf{w}_\alpha = \{w_i\}_{\alpha, i=1, \ldots, D} = \mathcal{C}\{u_i^\alpha\} \tag{8}$$

and Tsagris et al. (2011) defined the α-transformation, based on Equation (8), as

$$\mathbf{z}_\alpha = \frac{1}{\alpha}\mathbf{H}\left(D\mathbf{w}_\alpha - \mathbf{1}_D\right), \tag{9}$$

where $\mathbf{H}$ is the Helmert sub-matrix and $\mathbf{1}_D$ is the D-dimensional vector of 1s. While the power transformed vector $\mathbf{w}_\alpha$ in Equation (8) remains in the simplex $\mathbb{S}^{D-1}$, $\mathbf{z}_\alpha$ in Equation (9) is mapped onto a subset of $\mathbb{R}^{D-1}$. Furthermore, as $\alpha \to 0$, Equation (9) converges to the ilr transformation in Equation (6) (Tsagris et al., 2016). For convenience purposes, α is generally taken to be between -1 and 1, but when zeros occur in the data, α must be restricted to be strictly positive. The inverse of $\mathbf{z}_\alpha$ is

$$\mathbf{u}_\alpha = \mathcal{C}\left\{\left(\alpha \mathbf{H}^\top \mathbf{z}_\alpha + \mathbf{1}_D\right)^{1/\alpha}\right\}. \tag{10}$$

Tsagris et al. (2011) argued that while the α-transformation did not satisfy some of the properties that Aitchison (2003) deemed important, this was not a downside of this transformation as those properties were suggested mainly to fortify the concept of log-ratio methods. Scealy and Welsh (2014) also questioned the importance of these properties and, in fact, showed that some of them are not actually satisfied by the log-ratio methods that they were intended to justify.

2.2 Regression models for compositional data

2.2.1 Additive and isometric log-ratio regression models

Let $\mathbf{V}$ denote the response matrix with n rows containing alr transformed compositions. $\mathbf{V}$ can then be linked to some predictor variables $\mathbf{X}$ via

$$\mathbf{V} = \mathbf{XB} + \mathbf{E}, \tag{11}$$

where $\mathbf{B} = (\boldsymbol{\beta}_2, \ldots, \boldsymbol{\beta}_D)$ is the matrix of coefficients, $\mathbf{X}$ is the design matrix containing the p predictor variables (and the column of ones) and $\mathbf{E}$ is the residual matrix. Referring to Equation (2), Equation (11) can be re-written as

$$\log\left(\frac{u_{ij}}{u_{1j}}\right) = \mathbf{X}_j^\top \boldsymbol{\beta}_i \Leftrightarrow \log u_{ij} = \log u_{1j} + \mathbf{X}_j^\top \boldsymbol{\beta}_i, \quad i = 2, \ldots, D, \; j = 1, \ldots, n, \tag{12}$$

where $\mathbf{X}_j^\top$ denotes the jth row of $\mathbf{X}$. Equation (12) can be found in Tsagris (2015b) where it is shown that the alr regression (11) is in fact a multivariate linear regression in the logarithm of the compositional data with the first component (or any other component) playing the role of an offset variable; an independent variable with coefficient equal to 1.

Regression based on the ilr tranformation (ilr regression) is similar to alr regression and is carried out by substituting $\mathbf{V}$ in Equation (11) by $\mathbf{Z}_0$ in Equation (6). The fitted values for both the alr and ilr transformations are the same and are therefore generally back transformed onto the simplex using the inverse of the alr transformation in Equation (3) for ease of interpretation.

2.2.2 Kullback-Leibler divergence based regression

Murteira and Ramalho (2016) estimated the $\boldsymbol{\beta}$ coefficients via minimization of the Kullback-Leibler divergence

$$\min_{\boldsymbol{\beta}} \sum_{j=1}^{n} \mathbf{u}_j^\top \log \frac{\mathbf{y}_j}{\hat{\boldsymbol{\mu}}_j} = \max_{\boldsymbol{\beta}} \sum_{j=1}^{n} \mathbf{u}_j^\top \log \hat{\boldsymbol{\mu}}_j, \tag{13}$$

where $\hat{\boldsymbol{\mu}}_j = (\hat{\mu}_{j1}, \ldots, \hat{\mu}_{jD})^T$ is

$$\hat{\mu}_{ij} = \left\{ \begin{array}{ll} \frac{1}{1+\sum_{k=2}^{D} e^{\mathbf{x}_j^T \boldsymbol{\beta}_k}} & \text{if} \;\; i = 1 \\ \frac{e^{\mathbf{x}_j^T \boldsymbol{\beta}_i}}{1+\sum_{k=2}^{D} e^{\mathbf{x}^T \boldsymbol{\beta}_k}} & \text{for} \;\; i = 2, ..., D, \end{array} \right\} \tag{14}$$

with $\boldsymbol{\beta}_i = (\beta_{0i}, \beta_{1i}, ..., \beta_{pi})^T$, $i = 1, \ldots, d$. (Tsagris, 2015a,b)

The regression model in Equation (13), also referred to as Multinomial logit regression, will be denoted by Kullback-Leibler Divergence (KLD) regression throughout the rest of the paper. KLD regression is a semi-parametric regression technique, and, unlike alr and ilr regression, can handle zeros naturally.

3 The α–k–NN and α–kernel Regression Models

Our proposed α–k–NN and α–kernel regression models extend the well-known k–NN regression and the Nadaraya-Watson smoother, respectively, to the compositional data setting. Unlike

previous approaches, the models are more flexible (due to the α parameter), entirely non-parametric and allow for zeros, thus filling an important gap in the compositional data analysis literature

In general terms, to predict the response value corresponding to a new vector of predictor values ($\mathbf{x}_{new}$), the k–NN algorithm first computes the Euclidean distances from $\mathbf{x}_{new}$ to the observed predictor values $\mathbf{x}$. k–NN regression works by selecting the response values coinciding with the observations with the k–smallest distances between $\mathbf{x}_{new}$ and the observed predictor values, and then averaging those response values, using the sample mean or median, for example. The proposed α–k–NN regression algorithm relies on an extension of the sample mean to the sample Fréchet mean, described below, while the α–kernel regression method further extends this mean.

3.1 The Fréchet mean

The Fréchet mean (Pennec, 1999) on a metric space ($\mathbb{M}$, dist), with the distance between $p, q \in \mathbb{M}$ given by dist (p, q), is defined by $\operatorname{argmin}_{h \in \mathbb{M}} E_U \left[\operatorname{dist}(U, h)^2 \right]$ and $\operatorname{argmin}_{h \in \mathbb{M}} \sum_{j=1}^{n} \operatorname{dist}(u_j, h)^2$ in the population and finite sample cases, respectively. The Fréchet mean in the compositional data setting for a sample size n is the argument $\boldsymbol{\gamma} = (\gamma_1, \ldots, \gamma_D)^T$ which minimizes the following expression in Tsagris et al. (2011)

$$\boldsymbol{\mu}_\alpha = \sum_{j=1}^{n} \left[\sum_{i=1}^{D} \left(\frac{u_{ij}^\alpha}{\sum_{k=1}^{D} u_{kj}^\alpha} - \gamma_i \right)^2 \right],$$

where $\gamma_i = \frac{\mu_i^\alpha}{\sum_{k=1}^{D} \mu_k^\alpha}$. Setting $\partial \boldsymbol{\mu}_\alpha / \partial \gamma_i = 0$, the minimization occurs at

$$\hat{\gamma}_i = \frac{1}{n} \sum_{j=1}^{n} \left(\frac{u_{ij}^\alpha}{\sum_{k=1}^{D} u_{kj}^\alpha} \right),$$

and thus

$$\hat{\boldsymbol{\mu}}_\alpha(\mathbf{u}) = \mathcal{C} \left\{ \left\{ \left[\frac{1}{n} \sum_{j=1}^{n} \left(\frac{u_{ij}^\alpha}{\sum_{k=1}^{D} u_{kj}^\alpha} \right) \right]^{1/\alpha} \right\}_{i=1,\ldots,D} \right\} = \mathcal{C} \left\{ \left\{ \left(\sum_{j=1}^{n} u_{ij}^\alpha \right)^{1/\alpha} \right\}_{i=1,\ldots,D} \right\}. \quad (15)$$

Kendall and Le (2011) showed that the central limit theorem applies to Fréchet means defined on manifold valued data and the simplex space is an example of a manifold (Pantazis et al., 2019). Another nice property of the Fréchet mean is that it is absolutely continuous as α tends to zero, in the limiting case, Equation (15) converges to the closed geometric mean, $\hat{\boldsymbol{\mu}}_0$ defined below and in Aitchison (1989). That is,

$$\lim_{\alpha \to 0} \hat{\boldsymbol{\mu}}_\alpha(\mathbf{u}) \to \hat{\boldsymbol{\mu}}_0(\mathbf{u}) = \mathcal{C} \left\{ \left\{ \left(\prod_{j=1}^{n} u_{ij} \right)^{1/n} \right\}_{i=1,\ldots,D} \right\}.$$

3.2 The α–k–NN regression

When the response variables, $\mathbf{u}_1, \mathbf{u}_2, \ldots, \mathbf{u}_n$, represent compositional data, the k–NN algorithm can be applied to the transformed data in a straightforward manner, for a specified transformation appropriate for compositional data. For added flexibility and to be able to handle

compositional data with zeros directly, we propose extending k–NN regression using the power transformation in Equation (8) combined with the Fréchet mean in Equation (15). Specifically, in α–k–NN regression, the predicted response value corresponding to $\mathbf{x}_{new}$ is then

$$\hat{\boldsymbol{\mu}}_\alpha \left(\mathbf{u}_j | j \in \mathcal{A}\right) = \hat{\boldsymbol{\mu}}_{\alpha,new} = \mathcal{C} \left\{ \left\{ \left(\sum_{j \in \mathcal{A}} u_{ij}^\alpha \right)^{1/\alpha} \right\}_{i=1,\ldots,D} \right\}, \tag{16}$$

where $\mathcal{A}$ denotes the set of k observations, the k nearest neighbours. In the limiting case of $\alpha = 0$, the predicted response value is then

$$\hat{\boldsymbol{\mu}}_0 \left(\mathbf{u}_j | j \in \mathcal{A}\right) = \hat{\boldsymbol{\mu}}_{new} = \mathcal{C} \left\{ \left\{ \left(\prod_{j \in \mathcal{A}} u_{ij} \right)^{1/k} \right\}_{i=1,\ldots,D} \right\}.$$

It is interesting to note that the limiting case also results from applying the clr transformation to the response data, taking the mean of the relevant k transformed observations and then back transforming the mean using Equation (5). To see this, let

$$\hat{\mathbf{y}}_{new} = \left\{ \frac{1}{k} \sum_{j \in \mathcal{A}} y_{ij} \right\}_{i=1,\ldots,D} = \left\{ \frac{1}{k} \sum_{j \in \mathcal{A}} \log \frac{u_{ij}}{g\left(\mathbf{u}_j\right)} \right\}_{i=1,\ldots,D}$$

and hence

$$\hat{\mathbf{u}}_{new} = \mathcal{C} \left\{ \left\{ \frac{e^{\hat{y}_{new,i}}}{\sum_{l=1}^D e^{\hat{y}_{new,i}}} \right\}_{i=1,\ldots,D} \right\} = \mathcal{C} \left\{ \left\{ \frac{e^{\frac{1}{k} \sum_{j \in \mathcal{A}} \log \frac{u_{ij}}{g\left(\mathbf{u}_j\right)}}}{\sum_{l=1}^D e^{\frac{1}{k} \sum_{j \in \mathcal{A}} \log \frac{u_{jj}}{g\left(\mathbf{u}_j\right)}}} \right\}_{i=1,\ldots,D} \right\}$$

$$= \mathcal{C} \left\{ \left\{ \frac{\prod_{j \in \mathcal{A}} u_{ij}^{1/k}}{\sum_{l=1}^D \prod_{j \in \mathcal{A}} u_{lj}^{1/k}} \right\}_{i=1,\ldots,D} \right\} = \mathcal{C} \left\{ \left\{ \prod_{j \in \mathcal{A}} u_{ij}^{1/k} \right\}_{i=1,\ldots,D} \right\}.$$

A visual representation of the added flexibility provided by α on the Fréchet mean, in addition to the effect of k, is given in Figure 1. Figure 1 shows how the Fréchet mean varies when α moves from -1 to 1 and different nearest neighbours are used. In Figures 1(a) and 1(b) two different sets of 15 neighbours are used, whereas in Figure 1(c) all 25 nearest neighbours are used. The closed geometric mean (i.e. Fréchet mean with $\alpha = 0$) may lay outside the body of those observations. The Fréchet mean on the contrary offers a higher flexibility, which in conjunction with the number of nearest neighbour k yields estimates that can lie within the region of the selected set of compositional observations as visualised in Figure 1.

3.3   The α–kernel regression

Another possible non-parametric regression model results from extending the α regression estimator by slightly modifying Equation (16) and adopting the Nadaraya-Watson estimator (Nadaraya, 1964, Watson, 1964). This gives rise to the α–kernel regression based on the weighted mean as follows
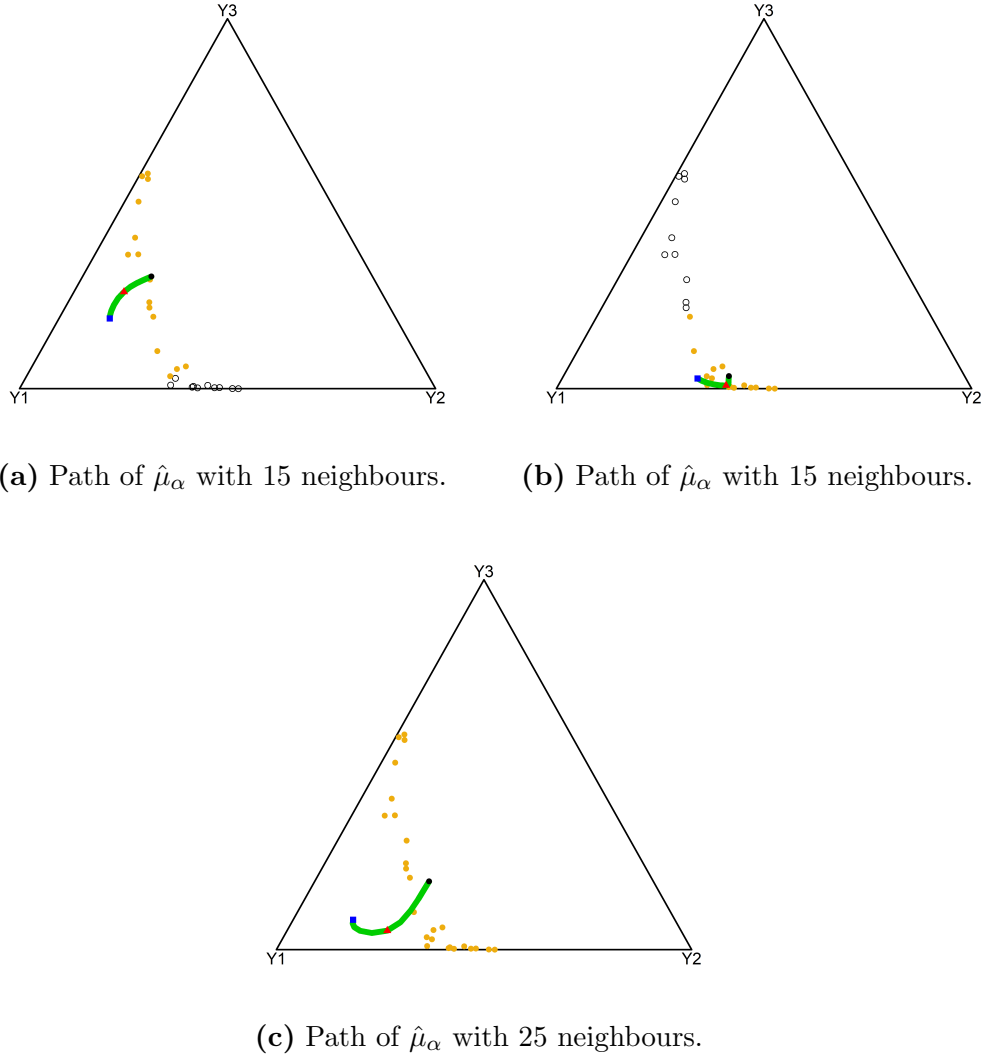
(a) Path of $\hat{\mu}_\alpha$ with 15 neighbours.

(b) Path of $\hat{\mu}_\alpha$ with 15 neighbours.

(c) Path of $\hat{\mu}_\alpha$ with 25 neighbours.

Figure 1: **Effect of α and k**. The symbols are as follows: $\blacksquare$ = Fréchet mean with $\alpha = -1$, $\blacktriangle$ = Fréchet mean with $\alpha = 0$ and $\bullet$ = Fréchet mean with $\alpha = 1$. The dashed green curve - shows the path of all Fréchet means starting with $\alpha = -1$ up to $\alpha = 1$. The golden circles indicate the set of observations used to compute the Fréchet mean.

$$
\begin{aligned}
\hat{\boldsymbol{\mu}}_{\alpha,new} &= \mathcal{C}\left\{\left\{\left(\frac{\sum_{j=1}^{n} K_h(\mathbf{x}_j - \mathbf{x}_{new})u_{ij}^\alpha}{\sum_{j=1}^{n} K_j(\mathbf{x}_j - \mathbf{x}_{new})}\right)^{1/\alpha}\right\}_{i=1,\ldots,D}\right\} \\
&= \mathcal{C}\left\{\left\{\left(\sum_{j=1}^{n} K_h(\mathbf{x}_j - \mathbf{x}_{new})u_{ij}^\alpha\right)^{1/\alpha}\right\}_{i=1,\ldots,D}\right\},
\end{aligned}
\tag{17}
$$

where $\mathbf{x}_j$ and $\mathbf{x}_{new}$ denote the j-th row vector and the new row vector containing the p predictor variables, and $K_h(\mathbf{x} - \mathbf{y})$ denotes the kernel function to be employed. Typical examples include the Gaussian (or radial basis function) $K_h(\mathbf{x} - \mathbf{y}) = e^{-\frac{\|\mathbf{x}-\mathbf{y}\|^2}{2h^2}}$, or the Laplacian kernel $K_h(\mathbf{x} - \mathbf{y}) = e^{-\frac{\|\mathbf{x}-\mathbf{y}\|}{h}}$, where $\|\mathbf{x} - \mathbf{y}\|$ denotes the Euclidean distance between row vectors $\mathbf{x}$ and $\mathbf{y}$.

The α–k–NN regression is more flexible than the simple k–NN regression but it is a special case of α–kernel regression and thus less flexible. The difference lies in the fact that the α–kernel regression produces a weighted Fréchet mean computed using all observations whereas the α–k–NN regression computes the Fréchet mean using a subset of the observations. However, the advantage of the α–k–NN regression over the α–kernel regression is the former's high computational efficiency. Lastly, as $\alpha \to 0$ the α–kernel regression (17) converges to the Nadaraya-Watson estimator using the log-ratio transformed data (Di Marzio et al., 2015). This is straightforward to show by substituting the u_{ij}^{α} with the isometric α-transformation (9).

3.4 Theoretical remarks

The general family of nearest regression estimators, under weak regularity conditions, were shown to be uniformly consistent with probability one and the corresponding rate of convergence is near-optimal (Cheng, 1984). More recently, Jiang (2019) proved the non-asymptotic uniform rates of consistency for the k–NN regression[2]. The proof of consistency of the α–k–NN regression estimator falls within the work of Lian et al. (2011) who dealt with the case of the response variable belonging in a separable Hilbert space. Lian et al. (2011) investigated the rates of strong (almost sure) convergence of the k–NN estimate under finite moment conditions and exponential tail condition on the noises. Recall that the simplex in Equation (1) is a $D-1$ Hilbert space (Pawlowsky-Glahn and Egozcue, 2001). However, results (asymptotic properties) for the α–k–NN and α–kernel regressions are much harder to derive due to the introduction of the power parameter α and are not considered here.

3.5 Cross-Validation protocol to select the values of α and k or h

The 10-fold Cross-Validation (CV) protocol is utilised to tune the pair (α, k) or (α, h). In the 10-fold CV pipeline, the data are randomly split into 10 folds of nearly equal sizes. One fold is selected to play the role of the test set, while the other folds are considered the training set. The regression models are fitted on the training set and their predictive capabilities are estimated using the test set. This procedure is repeated for the 10 folds so that each fold plays the role of the test set. Ultimately, the predictive performance of each regression model is computed from the aggregation of their predictive performances at each fold. Note that while for Euclidean data the criterion of predictive performance is typically the mean squared error, we instead measure the Kulback-Leibler (KL) divergence from the observed to the predicted compositional vectors as well as the Jensen-Shannon (JS) divergence which, unlike KL, is a metric, to account for the compositional nature of our response data. The KL and JS measures of divergence are given below:

$$\text{KL}(\mathbf{y}, \hat{\mathbf{y}}) \;=\; \sum_{i=1}^{D} y_i \log \frac{y_i}{\hat{y}}. \tag{18a}$$

$$\text{JS}(\mathbf{y}, \hat{\mathbf{y}}) \;=\; \sum_{i=1}^{D} \left(y_i \log \frac{2y_i}{y_i + \hat{y}_i} + \hat{y}_i \log \frac{2\hat{y}_i}{y_i + \hat{y}_i} \right). \tag{18b}$$

---

[2]For more asymptotic results, see the references cited within Jiang (2019).

4 Simulation studies

Monte Carlo simulation studies were implemented to assess the predictive performance of the proposed α–k–NN and α–kernel regressions compared to the KLD regression, an alternative semi-parametric approach that also allows for zeros. While KLD regression is computationally expensive, employment of the Newton-Raphson algorithm (Böhning, 1992) allows for efficient and feasible computation.

Multiple types of relationships between the response and predictor variables are considered and a 10-fold CV protocol was applied for each regression model to evaluate its predictive performance. A second axis of comparison was an assessment of computational cost of the regressions. Using the same scenario as above, the computational efficiency of the α–k–NN regression was compared to that of the KLD regression.

All computations were carried out on a laptop with Intel Core i5-5300U CPU at 2.3GHz with 16 GB RAM and SSD installed using the R package *Compositional* (Tsagris et al., 2022) for all regression models.

4.1 Predictive performance

In our simulation study, the values of the one or two predictor variables (denoted by $\mathbf{x}$) were generated from a Gaussian distribution with mean zero and unit variance, and were linked to the compositional responses via two functions: a polynomial as well as a more complex segmented function. For both cases, the outcome was mapped onto $\mathbb{S}^{D-1}$ using Equation (19)

$$y_i = \left\{ \begin{array}{ll} \frac{1}{1+\sum_{k=2}^{D} e^{f_k(\mathbf{x})}} & i = 1 \\ \frac{e^{f_i(\mathbf{x})}}{1+\sum_{k=2}^{D} e^{f_k(\mathbf{x})}}, & i = 2, ..., D \end{array} \right\}. \tag{19}$$

More specifically, for the simpler polynomial case, the values of the predictor variables were raised to a power (1, 2 or 3) and then multiplied by a vector of coefficients. White noise ($\mathbf{e}_j$) was added as follows

$$f_i\left(\mathbf{x}_j\right) = \left(\mathbf{x}_j^{\nu}\right)^T \boldsymbol{\beta}_i + \mathbf{e}_j, \ i = 2, \ldots, D, \ j = 1, \ldots, n \tag{20}$$

where $\nu = 1, 2, 3$ indicates the degree of the polynomial. The constant terms in the regression coefficients $\boldsymbol{\beta}_i, \ i = 2, \ldots, D$ were randomly generated from $N(-3, 1)$ whereas the slope coefficients were generated from $N(2, 0.5)$.

For the segmented linear model case, one predictor variable x was set to range from -1 up to 1 and the f function was defined as

$$f_i\left(\mathbf{x}_j\right) = \left\{ \begin{array}{ll} x_j^2 \beta_{1i} + e_j, & \text{when } x_j > 0 \\ x_j^3 \beta_{2i} + e_j, & \text{when } x_j < 0, \end{array} \right. \tag{21}$$

where $i = 2, \ldots, D$ and $j = 1, \ldots, n$. The regression coefficients β_{1i} were randomly generated from a $N(-1, 0.3)$ while the regression coefficients $\beta_{2i}, \ i = 2, \ldots, D$, were randomly generated from $N(1, 0.2)$.

The above two scenarios were repeated with the addition of zero values in 20% of randomly selected compositional vectors. For each compositional vector that was randomly selected, a

third of its component values were set to zero and those vectors were normalised to sum to 1. Finally, for all cases, the sample sizes varied between 100 and 1,000 with an increasing step size equal to 50 while the number of components was set equal to $D = \{3, 5, 7, 10\}$. The estimated predictive performance of the regression models was computed using the KL divergence (18a) and the JS divergence (18b). The results, however, were similar so only the KL divergence results are shown. For all examined case scenarios the results were averaged over 100 repeats.

Figures 2 and 3 show graphically the results of the comparison of α–k–NN and α–kernel regressions with KLD regression with no zeros and zero values present, respectively. Note that values below 1 indicate that the proposed regression models have smaller prediction errors than the KLD regression. For the first case of no zero values present (Figure 2), when the relationship between the predictor variable(s) and the compositional responses is linear ($\nu = 1$), the error is slightly less for KLD regression compared to α–k–NN regression. The relative error is smaller between the α–kernel and the KLD regressions. In all other cases (quadratic, cubic and segmented relationships), the α–k–NN and the α–kernel regressions consistently produce more accurate predictions. Another characteristic observable in all plots of Figure 2 is that the relative predictive performance of both non-parametric regressions compared to KLD regression reduces as the number of components in the compositional data increases.
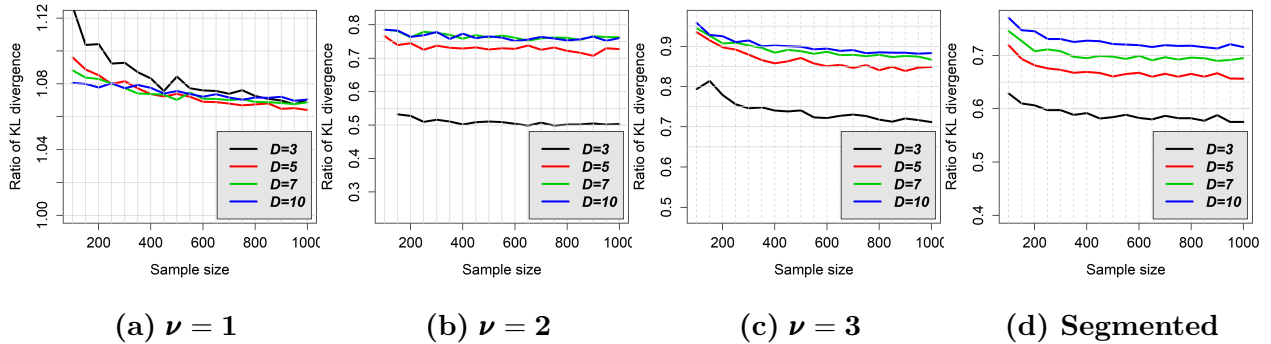
The results in the zero values present case in Figure 3 are, in essence, the same compared to the previous case. When the relationship between the predictor variables is linear, KLD regression again exhibits slightly more accurate predictions compared to the α–k–NN and the α–kernel regressions, while the opposite is true for most other cases. Furthermore, the impact of the number of components of the compositional responses on the relative predictive performance of both non-parametric regressions compared to the KLD regression varies according to the relationship between the response and covariates, the number of predictor variables and the sample size.

4.2 Computational efficiency of the α–k–NN regression

The linear relationship scenario (that is, when the degree of the polynomial in Equation (20) is equal to $\nu = 1$), but without zero values was used to illustrate the computational efficiency of only the α–k–NN regression. Unlike the α–kernel regression, the α–k–NN regression does not require all observations for computation of the predicted values, only the k-nearest neighbours. Therefore,the α–kernel regression is computationally slower.

With massive data (tens or hundreds of thousands of observations) the search for the k nearest neighbours in the α–k–NN regression takes place using kd-trees implemented in the R package $RANN$ (Arya et al., 2019). An important advantage of a kd-tree is that it runs in $O(n \log n)$ time, where n is the sample size of the training data set. The $RANN$ package utilizes the Approximate Near Neighbor (ANN) C++ library, which can give the exact near neighbours or (as the name suggests) approximate near neighbours to within a specified error bound, but when the error bound is 0 (as in this case) an exact nearest neighbour search is performed. When the sample sizes are at the order of a few tens of thousands or less, kd-trees are slower to use and, in this case, the α–k–NN regression algorithm employs a C++ implemented function to search for the k nearest neighbours from the R package $Rfast$ (Papadakis et al., 2022). In

α–k–NN regression

(a) ν = 1   (b) ν = 2   (c) ν = 3   (d) Segmented

α–kernel regression

(e) ν = 1   (f) ν = 2   (g) ν = 3   (h) Segmented

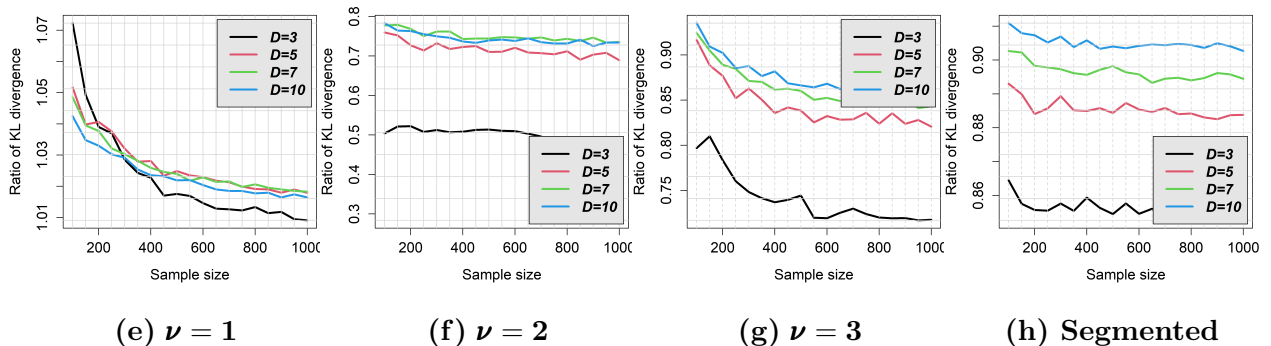Figure 2: **No zero values present case scenario**. Ratio of the Kullback-Leibler divergences between the α–k–NN and the KLD regression and between the α–kernel and the KLD regression. The number of components (D) appear with different colors. Values less than 1 indicate that the α–k–NN or the α–kernel regression has smaller prediction error than the KLD regression. (a) and (e): The degree of the polynomial in (20) is $\nu = 1$, (b) and (f): The degree of the polynomial in (20) is $\nu = 2$, (c) and (g): The degree of the polynomial in (20) is $\nu = 3$. (d) and (h) refer to the segmented linear relationship case (21).
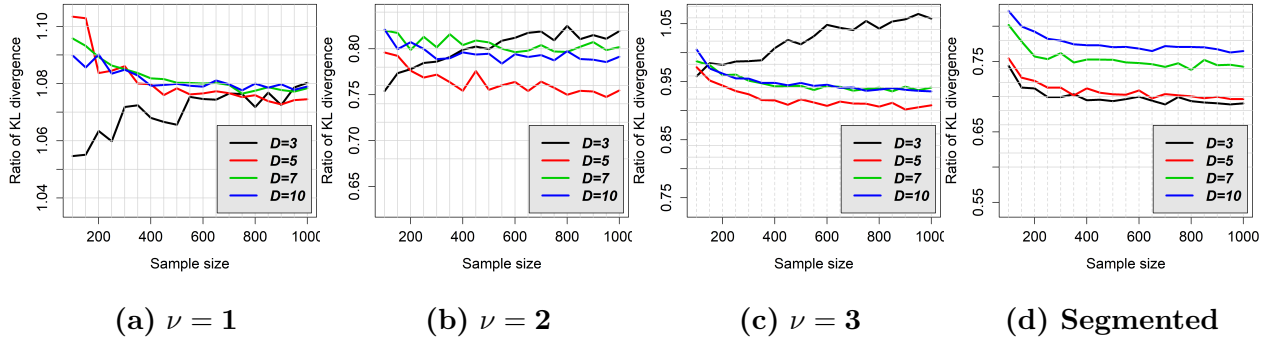
either case, with low or large sample sizes, it is worthy to highlight that the implementation of the α–k–NN regression algorithm has low memory requirements.

The KLD regression has been implemented computationally efficiently in R using the Newton-Raphson algorithm, yet it will be shown not to be as computational efficient as the α–k–NN regression.

In this study, the number of components of the compositional data and the number of predictor variables were kept the same as before but the sample sizes, the values of α and the number of neighbours were modified. Specifically, large sample sizes ranging from $500,000$ up to $10,000,000$ with an increasing step equal to $500,000$ were generated. Eleven positive values of α $(0, 0.1, \ldots, 1)$ were used and a large sequence of neighbours (from $k = 2$ up to $k = 100$ neighbours) were considered. The computational efficiency of each regression was measured as the time required to predict the compositional responses of $1,000$ new values.

The average time (in seconds) based on 10 repetitions versus the sample size of each re-

α–k–NN regression

(a) $\nu = 1$ **(b)** $\nu = 2$ **(c)** $\nu = 3$ **(d) Segmented**

α–kernel regression

(e) $\nu = 1$ **(f)** $\nu = 2$ **(g)** $\nu = 3$ **(h) Segmented**
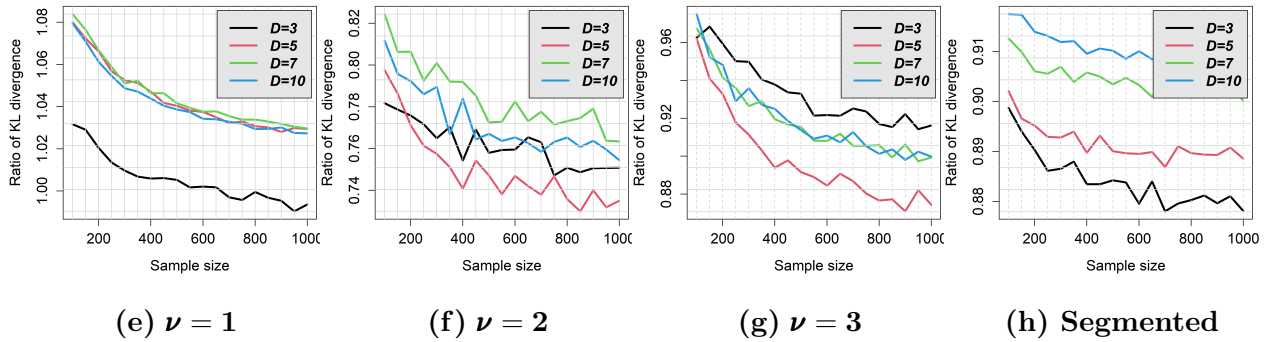
Figure 3: **Zero values present case scenario**. Ratio of the Kullback-Leibler divergences between the α–k–NN and the KLD regression and between the α–kernel and the KLD regression. The number of components (D) appear with different colors. Values less than 1 indicate that the α–k–NN or the α–kernel regression has a smaller prediction error than the KLD regression. (a) and (e): The degree of the polynomial in (20) is $\nu = 1$, (b) and (f): The degree of the polynomial in (20) is $\nu = 2$, (c) and (g): The degree of the polynomial in (20) is $\nu = 3$. (d) and (h) refer to the segmented linear relationship case (21).

gression is presented in Tables 1 and 2. The scalability of α–k–NN regression is better than that of KLD regression and as the sample size explodes the difference in the computational cost increases. Furthermore, the ratio of the computational cost of α–k–NN regression to the cost of OLS decays with the sample size and with the number of components. The converse is true for the ratio of the computational cost of KLD to the cost of OLS with respect to the number of components which appears to increase with the number of components. To appreciate the level of computational difficulty it should be highlighted that the α–k–NN regression produced a collection of $11 \times 99 = 1089$ predicted compositional data sets (for each combination of α and k). KLD, in contrast, produced a single set of predicted compositional data. The same is true for OLS regression whose time required for the same task is also presented.

Table 1: Computational cost of the regression with no zero values present. Time in seconds, of the each regression method for various sample size required by each regression for various sample sizes and $D = 3$ and $D = 5$ components. Inside the parentheses appear the ratios of the computational cost of each method compared to the computational cost of OLS.

	$D = 3$			$D = 5$		
Sample size	OLS	α–k–NN	KLD	OLS	α–k–NN	KLD
$n = 1 \times 10^6$	0.18	3.13(17.10)	3.17(17.34)	0.32	3.99(12.40)	7.77(24.14)
$n = 2 \times 10^6$	0.40	4.86(12.15)	7.34(18.36)	0.78	4.89(6.29)	14.57(18.75)
$n = 3 \times 10^6$	0.64	7.91(12.40)	12.72(19.94)	1.14	6.46(5.65)	21.70(18.99)
$n = 4 \times 10^6$	0.80	10.93(13.57)	16.84(20.92)	1.29	8.60(6.65)	31.20(24.17)
$n = 5 \times 10^6$	1.35	11.65(8.62)	17.49(12.94)	1.58	13.13(8.30)	43.71(27.63)
$n = 6 \times 10^6$	1.68	15.87(9.46)	22.44(13.38)	2.25	12.34(5.48)	43.25(19.21)
$n = 7 \times 10^6$	2.14	17.91(8.38)	27.23(12.75)	2.19	15.90(7.24)	47.83(21.79)
$n = 8 \times 10^6$	2.10	21.92(10.45)	28.76(13.71)	2.54	17.71(6.98)	60.49(23.83)
$n = 9 \times 10^6$	2.34	19.18(8.18)	29.34(12.52)	2.94	20.77(7.06)	70.96(24.13)
$n = 10 \times 10^6$	2.45	19.79(8.07)	31.92(13.01)	3.28	21.47(6.55)	80.02(24.41)

Table 2: Computational cost of the regression with no zero values present. Time in seconds, of the each regression method versus the sample size required by each regression for various sample sizes and $D = 7$ and $D = 10$ components. Inside the parentheses appear the ratios of the computational cost of each method compared to the computational cost of OLS.

	$D = 7$			$D = 10$		
Sample size	OLS	α–k–NN	KLD	OLS	α–k–NN	KLD
$n = 1 \times 10^6$	0.50	4.11(8.19)	12.06(24.02)	0.57	4.78(8.38)	20.98(36.81)
$n = 2 \times 10^6$	0.91	5.51(6.03)	24.35(26.64)	1.28	6.74(5.28)	48.88(38.30)
$n = 3 \times 10^6$	1.47	7.57(5.15)	36.65(24.95)	2.06	9.58(4.65)	76.62(37.19)
$n = 4 \times 10^6$	1.72	8.11(4.71)	44.51(25.83)	2.19	8.23(3.75)	78.63(35.86)
$n = 5 \times 10^6$	2.11	8.94(4.24)	53.77(25.50)	3.02	10.13(3.35)	102.63(33.93)
$n = 6 \times 10^6$	3.13	11.17(3.57)	65.74(21.00)	3.54	12.91(3.64)	133.86(37.78)
$n = 7 \times 10^6$	3.44	14.46(4.20)	82.31(23.90)	4.40	15.20(3.46)	171.33(38.95)
$n = 8 \times 10^6$	3.53	18.03(5.11)	108.84(30.87)	5.15	17.31(3.36)	199.97(38.84)
$n = 9 \times 10^6$	4.13	21.37(5.18)	117.37(28.43)	7.77	23.61(3.04)	263.47(33.92)
$n = 10 \times 10^6$	4.90	23.56(4.80)	139.84(28.51)	8.00	24.34(3.04)	312.00(38.98)

5 Examples with real data

5.1 Small sample sized data sets

To assess the predictive performance of α–k–NN and α–kernel regression in practice, 7 publicly available small sample sized data sets, with compositional responses, were utilised as examples. The same 10-fold CV protocol as before (see Subsection 3.5) was repeated using the 7 real data sets which are described briefly below. Note that the names of the data sets are consistent with the names previously used in the literature. A summary of the characteristics of the data sets, including the dimension of the response matrix, the number of compositional response vectors containing at least one zero value as well as the number of predictor variables, is provided in Table 3.

- **Lake**: Measurements in silt, sand and clay were taken at 39 different water depths in an Arctic lake. The question of interest was to predict the composition of these three elements for a given water depth. The data set is available in the R package *compositions* (van den Boogaart et al., 2018) and contains no zero values.

- **Glacial**: In a pebble analysis of glacial tills, the percentages by weight in 92 observations of pebbles of glacial tills sorted into 4 categories (red sandstone, gray sandstone, crystalline and miscellaneous) were recorded. The glaciologist was interested in predicting the compositions based on the total pebbles counts. The data set is available in the R package *compositions* (van den Boogaart et al., 2018) and almost half of the observations (42 out of 92) contain at least one zero value.

- **GDP**: The 2009 GDP per capita of the 27 member states of the European Union as well as the mean household consumption expenditures (in Euros) in 12 categories (food, house, alcohol-tobacco, clothing, household, health, transport, communication, culture, education, horeca and miscellaneous) are provided by eurostat. The data are available in Egozcue et al. (2012) and they contain no zero values.

- **Gemas**: This data set contains 2083 compositional vectors containing the concentration in 22 chemical elements (in mg/kg). The data set is available in the R package *robCompositions* (Templ et al., 2011) with 2108 vectors, but 25 vectors had missing values and thus were excluded from the current analysis. There was only one vector with one zero value. The predictor variables are the annual mean temperature and annual mean precipitation.

- **Fish**: This data set provides information on the mass (the only predictor variable) and 26 morphometric measurements (the compositional response data) for 75 Salvelinus alpinus (a type of fish). The data set is available in the R package *easyCODA* (Greenacre, 2018) and contains no zero values.

- **Data**: In this data set, the compositional response is a matrix of 9 party vote-shares across 89 different democracies (countries) and the (only) predictor variable is the average number of electoral districts in each country. The data set is available in the R package *compositions* (Rozenas, 2015) and 80 out of the 89 vectors contain at least one zero value.

- **Elections**: The Elections data set contains information on the 2000 U.S. presidential election in the 67 counties of Florida. The number of votes each of the 10 candidates received was transformed into proportions. For each county, information on 8 predictor variables was available such as population, percentage of population over 65 years old, mean personal income, percentage of people who graduated from college prior to 1990 among others. The data set is available in Smith (2002) and 23 out of the 67 vectors contained at least one zero value.

Figure 4 presents the boxplots of the relative performance (computed via the KL divergence) of α–k–NN and α–kernel regression compared to KLD regression for each data set. As before, values lower than 1 indicate that the proposed regression algorithm has smaller prediction error than KLD regression. On average, the α–k–NN outperformed the KLD regression for the data sets **GDP**, **Gemas** and **Elections**, whereas the opposite was true for the data set **Lake**, **Glacial**, **Fish** and **Data**. The conclusions are similar for the α–kernel regression and it had smaller prediction error (on average) for 4 of the 7 data sets, but this type of regression exhibited significantly less variability.

Table 3 presents the most frequently selected values of α and k. It is perhaps worth mentioning that the value of $\alpha = 0$ was never selected for any data set, indicating that the ilr transformation in Equation (6) was never considered the optimal transformation. When the percentage of times the value of α is selected is large, this implies a small variance in the chosen value of the parameter. From Table 3, it appears that the larger the value of the optimal α (roughly), the smaller the variance. For example, for the data set **Lake**, the optimal value α was chosen to be 1, 93% of the time for the α–k–NN regression and 95% of the time for the α–kernel regression. For the data set **Fish**, however, the optimal value of α, as well as the percentage of time it was chosen, were smaller for both regressions. There does not appear to be an association between the variability in the chosen value of k and the variability in the optimal α values for the α–k–NN regression. For example, for both data sets **Lake** and **Fish**, 10 nearest neighbours were chosen only 66% of the time. For the data set **Gemas**, the choice of α was highly variable, whereas the choice of k was always the same. The opposite was true for the data set **Data**, for which the optimal value of α was always the same but the value of k was highly variable.

The first data set (**Lake**) is of small enough dimension ($D = 3$) to be visualized. Figure 5 contains the observed compositional data, along with the fitted values of the α–k–NN regression and of the KLD regression, as well as the α–kernel regression and the KLD regression. The water depth increases as one moves from left to right showing the change in the three Arctic lake components. As expected, the fitted values of the proposed models do not fall along a line, as is the case with the KLD regression. In order to obtain a better picture, Figure 6 with the fitted values of each individual component versus the (logarithm of the) water depth. This figure better shows the individual fits for each component for each regression model indicating a better fit of the KLD at lower depths, i.e. closer to the surface of the lake.

Evidently, as the dimensions increase visual inspection of the fit of the models becomes harder. To bypass this problem the correlations between each component observed and fitted values may be computed. Figure 7 presents the boxplots of the correlations for each model
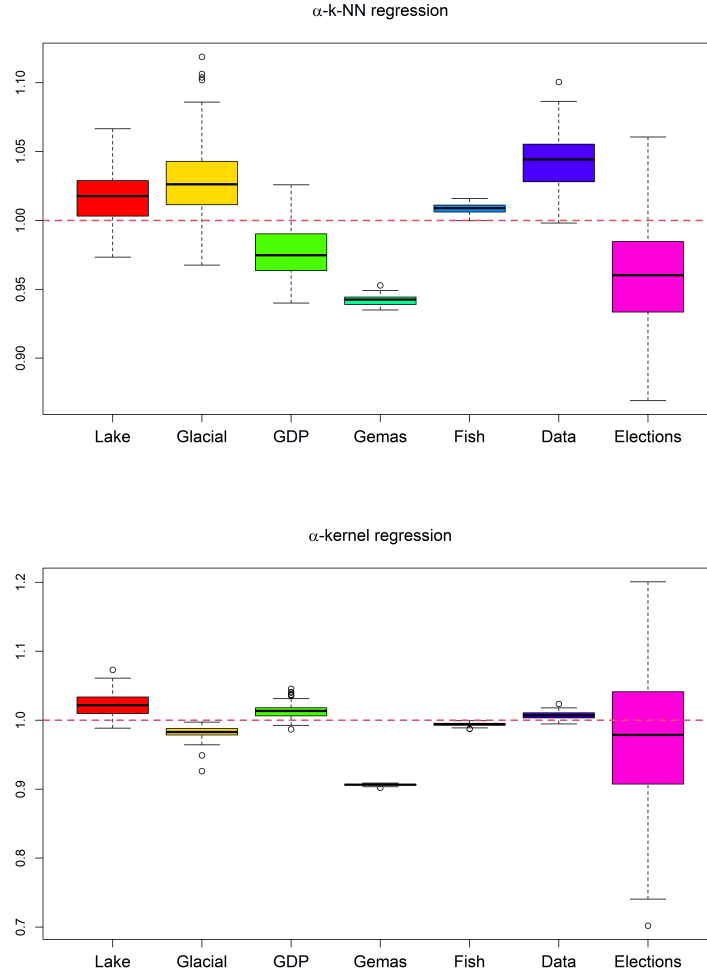
Figure 4: Ratio of the Kullback-Leibler divergence between the α–k–NN regression (top) and the α–kernel regression (bottom) to the KLD regression. Values lower than 1 indicate a smaller prediction error compared to the KLD regression.

under examination, where it is evident that the non-parametric models have superseded the KLD regression model.

The results from the real data analysis do not come by surprise and in fact corroborate the findings of the simulation studies. The simulation studies showed that when the relationship between the compositional responses and the independent variable(s) is linear, the α–k–NN and the α–kernel regression models do not offer an improved fit over their competitor, the KLD regression. But, when the relationship is not linear the non-parametric models are to be preferred. For the Arctic lake data for example, the correlations between the alr transformed compositional data and the logarithm of the water depth are high and a scatter plot clearly shows a highly linear relationship. For the Gemma data set though, the correlations between the alr transformed compositional data and the two independent variables are rather low. Hence, this could be used as a rule of thumb for initial inspection as to whether the KLD regression should be selected against a non-parametric regression model.

Table 3: Information about the response compositional variables, the predictor variables and the pairs (α, k) chosen for every data set, including the most frequently selected value and the percentage of times it is selected.

Data Set	Dimensions $(n \times D)$	No of vectors with zero values	No of predictors	α–k–NN % of times α was selected	% of times k was selected	α–kernel % of times α was selected
Lake	39×3	0	1	1 (93%)	10 (66%)	1 (95%)
Glacial	92×4	42	1	1 (93%)	10 (75%)	1 (48%)
GDP	27×12	0	1	0.8 (33%)	5 (89%)	0.6 (43%)
Gemas	2083×30	1	2	0.5 (47%)	10 (100%)	1 (100%)
Fish	75×26	0	1	-1 (29%)	20 (66%)	0.1 (9%)
Data	89×9	80	1	1 (100%)	10 (28%)	1 (100%)
Elections	67×10	23	8	0.5 (30%)	7 (32%)	0.1 (76%)

5.2 Large scale sample sized data sets

A benefit of α–k–NN regression is its high computational efficiency and hence we also illustrate its performance on two real large scale data sets. The α–kernel regression was not examined due to its high memory requirements.

- **Seoul pollution**: Air pollution measurement information in Seoul, South Korea, is provided by the Seoul Metropolitan Government 'Open Data Plaza'. This particular data set was downloaded from kaggle. The average values for 4 pollutants (SO_2, NO_2, CO, O_3) are available along with the coordinates (latitude and longitude) of each site. The data were normalised to sum to 1, so as to obtain the composition of each pollutant. In total, there are 639,073 of observations. Since the predictors in the **Seoul pollution** data set, the longitude and latitude, are expressed in polar coordinates they were first transformed to their Euclidean coordinates, using the R package *Directional* (Tsagris et al., 2022), in order to validly compute the Euclidean distances.

- **Electric power consumption**: This data set contains 1,454,154 measurements of electric power consumption in one household with a one-minute sampling rate over a period of almost 4 years. The measurements were gathered in a house located in Sceaux (7km of Paris, France) between December 2006 and November 2010 (47 months). Different electrical quantities and some sub-metering values are available. The data set is available to download from the UCI Machine Learning Repository. The response data comprised of 3 energy sub meterings measured in watt-hour of active energy: a) the kitchen, containing mainly a dishwasher, an oven and a microwave (hot plates are not electric but gas powered). b) the laundry room, containing a washing-machine, a tumble-drier, a refrigerator and a light and c) the electric water-heater and the air-conditioner. The data were again transformed to compositional data. There are 4 predictor variables, namely the global
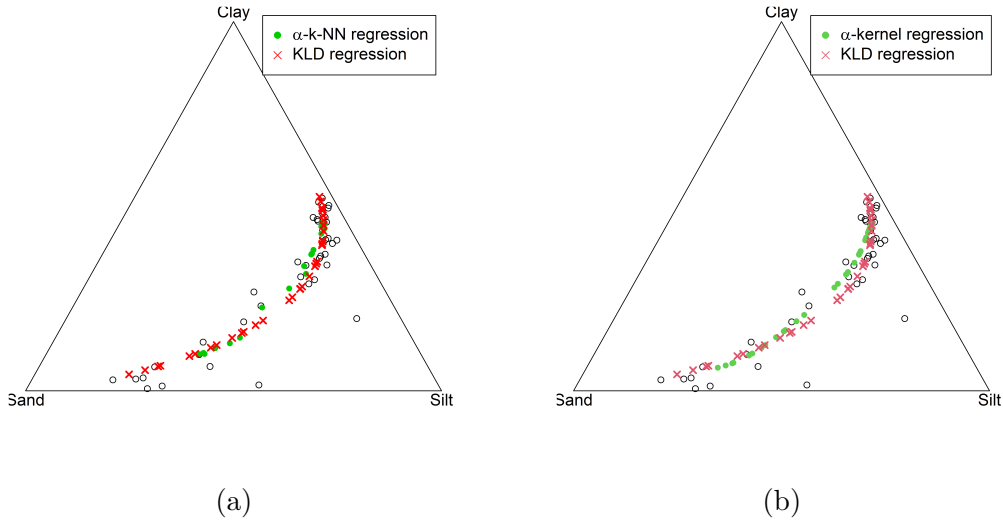
(a)   (b)

Figure 5: Ternary diagram of the Arctic lake data along with the fitted values based on the KLD regression and a) the α–k–NN regression and b) the α–kernel regression.

active power (in kilowatt), the global reactive power (in kilowatt), the voltage (in volt) and the global intensity (in ampere).

The same 10-fold CV protocol was employed again. Since both compositional data sets contained zero values, the KLD and α–k–NN regression methods were suitable. Only strictly positive values of α were therefore utilised, and for the nearest neighbours, 99 values were tested ($k = 2, \ldots, 100$). The results of the α–k–NN and the KLD regression are summarised in Table 4.

Table 4: The minimum Kullback-Leibler and Jensen-Shannon divergence obtained by the α–k–NN and the KLD regression for each large scale data. The optimal combination of α and k appears inside the parentheses.

Data Set	Kullbak-Leibler divergence		Jensen-Shannon divergence	
	α–k–NN	KLD	α–k–NN	KLD
Seoul Polution	0.068 (α=1, k=67)	0.047	0.005 (α=1, k=67)	0.004
Electric power consumption	0.541 (α=1, k=2)	0.825	0.046 (α=1, k=2)	0.063

These two large scale data sets suggest that, similarly to the small scale data sets, α–k–NN is a viable alternative regression model option for compositional data, with the advantage that it is as computationally efficient, or more than, other regression models, at the cost of interpretability of the effects of the predictor variables.
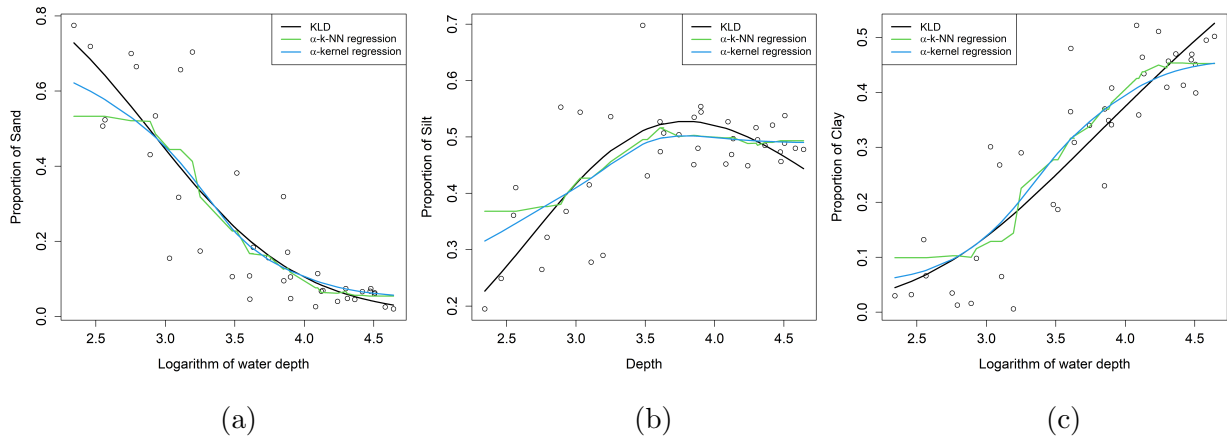
(a)   (b)   (c)

Figure 6: Arctic lake data: the logarithm of the water depth against the observed fitted values based on the three regression models.
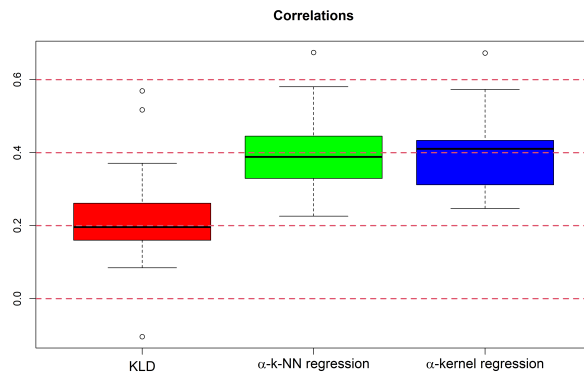

Figure 7: Arctic lake data: boxplot of the correlations between the observed and fitted values of each component for the three regression models.

6   Conclusions

Two generic regressions able to capture complex relationships involving compositional data, termed α–k–NN regression and α–kernel regression, were proposed that take into account the constraints on such data. Through simulation studies and the analysis of several real-life data sets, α–k–NN and α–kernel regressions were evaluated alongside a comparable, but semi-parametric, regression model available for this setting. The classical k–NN regression provided the foundation for α–k–NN regression, while the α–transformation was used to transform the compositional data. Using this transformation added to the flexibility of the model and meant that commonly occurring zero values in the compositional data were allowed, unlike with many other regression approaches for compositional data. The flexbility was further increased with the introduction of a weighted scheme computed via a Gaussian kernel. The Fréchet mean (defined for compositional data by Tsagris et al. (2011)) was used in order to prevent fitted values from being outside the simplex.

Using a CV procedure and pertinent measures of predictive performance, we found that in

simulation study cases where the relationship was non-linear (including when the data contained zeros), the α–k–NN and α–kernel regressions outperformed (sometimes substantially) their competing counterpart, KLD regression. For the real-life data sets, similar conclusions were made and our two non-parametric regressions tended to outperform KLD regression in data sets where it is surmised that non-linear relationships exist. We note that our conclusions were the same regardless of which type of divergence was used (either KL or JS). A second advantage of the α–k–NN regression solely, is its high computational efficiency as it can treat millions of observations in just a few seconds. We further highlight that the new regression techniques are publicly available in the R package *Compositional* (Tsagris et al., 2022).

A disadvantage of the α–k–NN regression, and of k–NN regression in general, is that it lacks the framework for classical statistical inference (such as hypothesis testing). This is counterbalanced by a) its higher predictive performance compared to parametric models and b) its high computational efficiency that make it applicable even with millions of observations. However, the use of ICE plots (Goldstein et al., 2015) that offer a visual inspection of the effect of each independent variable can overcome this issue. Note that while not considered in detail here, α-regression (Tsagris, 2015b) can also handle zeros but is computationally expensive[3].

With respect to possible future directions related to this work, one area is non-linear PCA, such as kernel PCA (Mika et al., 1999). However, a disadvantage of kernel methods is their computational cost that increases exponentially with sample size. Additionally, more variations and improvements of k–NN regression found in Nguyen et al. (2016) could be explored, which are applicable also in α–k–NN regression. Alternatively, the proposed regression models can be generalised towards random forests (Breiman, 2001), given the relationship of the latter with k–NN (Lin and Jeon, 2006), and potentially other machine learning techniques.

Instead of the α–transformation, one could use alternative transformations for compositional data and derive more transformation based k–NN regression estimators. The natural question is which one would perform best and under what circumstances. Further, the rationale of the α–k–NN regression can be transferred to Euclidean data by employing the Box-Cox transformation instead.

Projection pursuit (Friedman and Stuetzle, 1981) is another non-linear alternative to the α–k–NN regression, which also requires tuning of a parameter, but currently would only be possible for compositional data without zero values. While zero value imputation (Martín-Fernández et al., 2012, 2003) has the potential to resolve this problem, it could also lead to less accurate estimations as was found to be the case in (Tsagris, 2015a).

Conflict of interest

On behalf of all authors, the corresponding author states that there is no conflict of interest.

---

[3]A numerical optimization using the Nelder-Mead algorithm (Nelder and Mead, 1965) is applied using the command *optim* in R.

References

Aitchison, J. (1982). The statistical analysis of compositional data. *Journal of the Royal Statistical Society, Series B 44*(2), 139–177.

Aitchison, J. (1983). Principal component analysis of compositional data. *Biometrika 70*(1), 57–65.

Aitchison, J. (1989). Measures of location of compositional data sets. *Mathematical Geology 21*(7), 787–790.

Aitchison, J. (2003). *The statistical analysis of compositional data.* New Jersey: Reprinted by The Blackburn Press.

Arya, S., D. Mount, S. Kemp, and G. Jefferis (2019). *RANN: Fast Nearest Neighbour Search (Wraps ANN Library) Using L2 Metric.* R package version 2.6.1.

Bóhning, D. (1992). Multinomial logistic regression algorithm. *Annals of the Institute of Statistical Mathematics 44*(1), 197–200.

Breiman, L. (2001). Random forests. *Machine Learning 45*(1), 5–32.

Chen, E. Z. and H. Li (2016). A two-part mixed-effects model for analyzing longitudinal microbiome compositional data. *Bioinformatics 32*(17), 2611–2617.

Cheng, P. E. (1984). Strong consistency of nearest neighbor regression function estimators. *Journal of Multivariate Analysis 15*(1), 63–72.

Di Marzio, M., A. Panzera, and C. Venieri (2015). Non-parametric regression for compositional data. *Statistical Modelling 15*(2), 113–133.

Dryden, I. and K. Mardia (1998). *Statistical Shape Analysis.* John Wiley & Sons.

Egozcue, J., V. Pawlowsky-Glahn, G. Mateu-Figueras, and C. Barceló-Vidal (2003). Isometric logratio transformations for compositional data analysis. *Mathematical Geology 35*(3), 279–300.

Egozcue, J. J., J. Daunis-I-Estadella, V. Pawlowsky-Glahn, K. Hron, and P. Filzmoser (2012). Simplicial regression. the normal model. *Journal of Applied Probability and Statistics 6*(182), 87–108.

Friedman, J. H. and W. Stuetzle (1981). Projection pursuit regression. *Journal of the American Statistical Association 76*(376), 817–823.

Goldstein, A., A. Kapelner, J. Bleich, and E. Pitkin (2015). Peeking inside the black box: Visualizing statistical learning with plots of individual conditional expectation. *journal of Computational and Graphical Statistics 24*(1), 44–65.

Greenacre, M. (2018). *Compositional Data Analysis in Practice.* Chapman & Hall/CRC Press.

Gueorguieva, R., R. Rosenheck, and D. Zelterman (2008). Dirichlet component regression and its applications to psychiatric data. *Computational Statistics & Data Analysis 52*(12), 5344–5355.

Hijazi, R. and R. Jernigan (2009). Modelling compositional data using Dirichlet regression models. *Journal of Applied Probability and Statistics 4*(1), 77–91.

Iyengar, M. and D. K. Dey (2002). A semiparametric model for compositional data analysis in presence of covariates on the simplex. *Test 11*(2), 303–315.

Jiang, H. (2019). Non-Asymptotic Uniform Rates of Consistency for $k - NN$ Regression. In *Proceedings of the AAAI Conference on Artificial Intelligence*, Volume 33, pp. 3999–4006.

Katz, J. and G. King (1999). A statistical model for multiparty electoral data. *American Political Science Review 93*(1), 15–32.

Kendall, W. S. and H. Le (2011). Limit theorems for empirical fréchet means of independent and non-identically distributed manifold-valued random variables. *Brazilian Journal of Probability and Statistics 25*(3), 323–352.

Lancaster, H. (1965). The Helmert matrices. *American Mathematical Monthly 72*(1), 4–12.

Le, H. and C. Small (1999). Multidimensional scaling of simplex shapes. *Pattern Recognition 32*(9), 1601–1613.

Leininger, T. J., A. E. Gelfand, J. M. Allen, and J. A. Silander Jr (2013). Spatial Regression Modeling for Compositional Data With Many Zeros. *Journal of Agricultural, Biological, and Environmental Statistics 18*(3), 314–334.

Lian, H. et al. (2011). Convergence of functional k-nearest neighbor regression estimate with functional responses. *Electronic Journal of Statistics 5*, 31–40.

Lin, Y. and Y. Jeon (2006). Random forests and adaptive nearest neighbors. *Journal of the American Statistical Association 101*(474), 578–590.

Martín-Fernández, J., K. Hron, M. Templ, P. Filzmoser, and J. Palarea-Albaladejo (2012). Model-based replacement of rounded zeros in compositional data: Classical and robust approaches. *Computational Statistics & Data Analysis 56*(9), 2688–2704.

Martín-Fernández, J. A., C. Barceló-Vidal, and V. Pawlowsky-Glahn (2003). Dealing with zeros and missing values in compositional data sets using nonparametric imputation. *Mathematical Geology 35*(3), 253–278.

Melo, T. F., K. L. Vasconcellos, and A. J. Lemonte (2009). Some restriction tests in a new class of regression models for proportions. *Computational Statistics & Data Analysis 53*(12), 3972–3979.

Mika, S., B. Schölkopf, A. J. Smola, K.-R. Múller, M. Scholz, and G. Rátsch (1999). Kernel pca and de-noising in feature spaces. In *Advances in Neural Information Processing Systems*, pp. 536–542.

Morais, J., C. Thomas-Agnan, and M. Simioni (2018). Using compositional and Dirichlet models for market share regression. *Journal of Applied Statistics 45*(9), 1670–1689.

Mullahy, J. (2015). Multivariate fractional regression estimation of econometric share models. *Journal of Econometric Methods 4*(1), 71–100.

Murteira, J. M. R. and J. J. S. Ramalho (2016). Regression analysis of multivariate fractional data. *Econometric Reviews 35*(4), 515–552.

Nadaraya, E. A. (1964). On estimating regression. *Theory of Probability & Its Applications 9*(1), 141–142.

Nelder, J. and R. Mead (1965). A simplex algorithm for function minimization. *Computer Journal 7*(4), 308–313.

Nguyen, B., C. Morell, and B. De Baets (2016). Large-scale distance metric learning for k-nearest neighbors regression. *Neurocomputing 214*, 805–814.

Otero, N., R. Tolosana-Delgado, A. Soler, V. Pawlowsky-Glahn, and A. Canals (2005). Relative vs. absolute statistical analysis of compositions: a comparative study of surface waters of a mediterranean river. *Water Research 39*(7), 1404–1414.

Pantazis, Y., M. Tsagris, and A. T. Wood (2019). Gaussian asymptotic limits for the α-transformation in the Analysis of compositional data. *Sankhya A 81*(1), 63–82.

Papadakis, M., M. Tsagris, M. Dimitriadis, S. Fafalios, I. Tsamardinos, M. Fasiolo, G. Borboudakis, J. Burkardt, C. Zou, C. Lakiotaki, and C. Chatzipantsiou (2022). *Rfast: A Collection of Efficient and Extremely Fast R Functions*. R package version 2.0.6.

Pawlowsky-Glahn, V. and J. J. Egozcue (2001). Geometric approach to statistical analysis on the simplex. *Stochastic Environmental Research and Risk Assessment 15*(5), 384–398.

Pennec, X. (1999). Probabilities and statistics on riemannian manifolds: Basic tools for geometric measurements. In *IEEE Workshop on Nonlinear Signal and Image Processing*, Volume 4. Citeseer.

Rozenas, A. (2015). *ocomposition: Regression for Rank-Indexed Compositional Data*. R package version 1.1.

Scealy, J. and A. Welsh (2011). Regression for compositional data by using distributions defined on the hypersphere. *Journal of the Royal Statistical Society, Series B 73*(3), 351–375.

Scealy, J. and A. Welsh (2014). Colours and cocktails: Compositional data analysis 2013 Lancaster lecture. *Australian & New Zealand Journal of Statistics 56*(2), 145–169.

Shi, P., A. Zhang, and H. Li (2016). Regression analysis for microbiome compositional data. *The Annals of Applied Statistics 10*(2), 1019–1040.

Smith, R. L. (2002). A statistical assessment of Buchanan's vote in Palm Beach county. *Statistical Science 17*(4), 441–457.

Templ, M., K. Hron, and P. Filzmoser (2011). *robCompositions: an R-package for robust statistical analysis of compositional data.* John Wiley and Sons.

Tolosana-Delgado, R. and H. von Eynatten (2009). Grain-size control on petrographic composition of sediments: compositional regression and rounded zeros. *Mathematical Geosciences 41*(8), 869.

Tsagris, M. (2015a). A novel, divergence based, regression for compositional data. In *Proceedings of the 28th Panhellenic Statistics Conference, April, Athens, Greece.*

Tsagris, M. (2015b). Regression analysis with compositional data containing zero values. *Chilean Journal of Statistics 6*(2), 47–57.

Tsagris, M., G. Athineou, A. Alenazi, and C. Adam (2022). *Compositional: Compositional Data Analysis.* R package version 5.8.

Tsagris, M., G. Athineou, A. Sajib, E. Amson, M. Waldstein, and C. Adam (2022). *Directional: Directional Statistics.* R package version 5.5.

Tsagris, M., S. Preston, and A. Wood (2011). A data-based power transformation for compositional data. In *Proceedings of the 4rth Compositional Data Analysis Workshop, Girona, Spain.*

Tsagris, M., S. Preston, and A. T. Wood (2016). Improved classification for compositional data using the α-transformation. *Journal of Classification 33*(2), 243–261.

Tsagris, M. and C. Stewart (2018). A Dirichlet regression model for compositional data with zeros. *Lobachevskii Journal of Mathematics 39*(3), 398–412.

Tsagris, M. and C. Stewart (2020). A folded model for compositional data analysis. *Australian & New Zealand Journal of Statistics 62*(2), 249–277.

van den Boogaart, K., R. Tolosana-Delgado, and M. Bren (2018). *compositions: Compositional Data Analysis.* R package version 1.40-2.

Wand, M. P. and M. C. Jones (1994). *Kernel Smoothing.* Chapman and Hall/CRC.

Watson, G. S. (1964). Smooth regression analysis. *Sankhya: The Indian Journal of Statistics, Series A 26*(4), 359–372.

Xia, F., J. Chen, W. K. Fung, and H. Li (2013). A logistic normal multinomial regression model for microbiome compositional data analysis. *Biometrics 69*(4), 1053–1063.