

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

Michail Tsagris

Department of Economics, University of Crete, Rethymnon, Greece

mtsagris@uoc.gr

Co-authors: Connie Stewart (University of New Brunswick) and Abdulaziz Alenazi (Northern Border University)

DeptEcon 1/11/2023

What are Compositional Data?

Composition

A vector $\mathbf{x} = (x_1, x_2, \dots, x_D)$ denotes a D -part **composition** with sample space the simplex given by

$$\mathcal{S}^{D-1} = \left\{ \mathbf{x} = (x_1, x_2, \dots, x_D) : x_i > 0, i = 1, 2, \dots, D, \sum_{i=1}^D x_i = k \right\}$$

where k is a positive constant, usually 1 or 100.

Compositional Data Applications

- Compositional data arise in many different fields.
- Examples may be found in biology, ecology, economics, political science, forensic science, bioinformatics, etc...
- Many compositional data analysis applications involve covariates.
- The development of regression models for compositional data has been an active research area in recent years.

Applications involving compositional data and covariates can be classified according to the type of response and predictor data:

1. Compositional Response - Compositional Predictors
2. **Compositional Response - Euclidean Predictors**
3. Euclidean Response - Compositional Predictors

Arctic Lake Example: Compositional Response - Euclidean Predictor

- Samples from an Arctic lake ¹ were collected and the composition of sand, silt and clay were measured at different depths (in meters)
- $n = 39$, $D = 3$, $p = 1$

```
> library(compositions)
> data(ArcticLake)
> ArcticLake
  sand silt clay depth
1  77.5 19.5  3.0  10.4
2  71.9 24.9  3.2  11.7
3  50.7 36.1 13.2  12.8
4  52.2 40.9  6.6  13.0
5  70.0 26.5  3.5  15.7
```

¹[Aitchison, 2003]

1. Log-ratio (OLS) regression models²

- Let \mathbf{V} denote the response matrix with n rows containing log-ratio transformed compositions.
- The matrix \mathbf{V} is linked to predictor variables \mathbf{X} via

$$\mathbf{V} = \log \frac{\mathbf{Y}_{-1}}{Y_1} = \mathbf{XB} + \mathbf{E},$$

where

- $\mathbf{B} = (\beta_2, \dots, \beta_D)$ is the matrix of coefficients
 - \mathbf{X} is the design matrix
 - \mathbf{E} is the residual matrix.
- R package Compositional: `comp.reg()`.

²[Aitchison, 2003]

1. Log-ratio (OLS) regression models²

- Let \mathbf{V} denote the response matrix with n rows containing log-ratio transformed compositions.
- The matrix \mathbf{V} is linked to predictor variables \mathbf{X} via

$$\mathbf{V} = \log \frac{\mathbf{Y}_{-1}}{Y_1} = \mathbf{XB} + \mathbf{E},$$

where

- $\mathbf{B} = (\beta_2, \dots, \beta_D)$ is the matrix of coefficients
 - \mathbf{X} is the design matrix
 - \mathbf{E} is the residual matrix.
- R package Compositional: `comp.reg()`.

Zeros in the response data are problematic.

²[Aitchison, 2003]

2. Kullback-Leibler divergence (KLD) based regression³

- β coefficients are estimated via minimization of the Kullback-Leibler divergence

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

where the components of $\hat{\boldsymbol{\mu}}_j$ are given by

$$\hat{\mu}_i = \left\{ \begin{array}{ll} \frac{1}{1 + \sum_{k=1}^d e^{\mathbf{x}^{\top} \boldsymbol{\beta}_k}} & \text{if } i = 1 \\ \frac{e^{\mathbf{x}^{\top} \boldsymbol{\beta}_i}}{1 + \sum_{k=1}^d e^{\mathbf{x}^{\top} \boldsymbol{\beta}_k}} & \text{for } i = 2, \dots, D, \end{array} \right\}.$$

- R package Compositional: `kl.compreg()`.

Zeros in the response data are allowed.

³[Murteira and Ramalho, 2016]

- α - k - NN regression is an extension to the well-known k - NN regression.
- To predict a response value using k - NN regression:
 - Select response values corresponding to k observed predictor values that are *closest* to \mathbf{x}_{new} .
 - Average selected k response values.
- The proposed α - k - NN regression algorithm uses a flexible sample mean, appropriate for compositional data.

Fréchet mean⁴

For a sample of compositional data, the sample **Fréchet mean** is defined as

$$\hat{\boldsymbol{\mu}}_{\alpha}(\mathbf{y}) = \mathcal{C} \left\{ \left\{ \left(\sum_{j=1}^n y_{ji}^{\alpha} \right)^{1/\alpha} \right\}_{i=1, \dots, D} \right\},$$

where \mathcal{C} denotes the closure operation.

- 10-fold cross-validation (CV) is used to tune the pair (α, k) .
- *R* package *Compositional*: `aknnreg.tune()` and `aknn.reg()`.

Zeros in the response compositional data are allowed.

⁴[Tsagris et al., 2011]

- Kendall and Le⁵ 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 pantazis2019.
- A nice property of the Fréchet mean is that it is absolutely continuous as α tends to zero.
- In the limiting case (assuming strictly positive compositional data), the Fréchet mean converges to the closed geometric mean⁶

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

⁵[Kendall and Le, 2011]

⁶[Aitchison, 1989]

α -kernel Compositional-Euclidean Regression

For a sample of compositional data, the sample kernel weighted **Fréchet mean** is defined as

$$\hat{\boldsymbol{\mu}}_{\alpha}(\mathbf{y}) = \mathcal{C} \left\{ \left\{ \left(\sum_{j=1}^n K_h(\mathbf{x}_j - \mathbf{x}_{new}) y_{ij}^{\alpha} \right)^{1/\alpha} \right\}_{i=1, \dots, D} \right\},$$

where $K_h(\cdot)$ denotes the kernel function to be employed. Typical examples include the Gaussian (or radial basis function)

$K_h(\mathbf{x} - \mathbf{y}) = e^{-\frac{\|\mathbf{x} - \mathbf{y}\|_2^2}{2h^2}}$, or the Laplacian kernel $K_h(\mathbf{x} - \mathbf{y}) = e^{-\frac{\|\mathbf{x} - \mathbf{y}\|_1}{h}}$.

- 10-fold cross-validation (CV) is used to tune the pair (α, h) .
- R package *Compositional*: `akernreg.tune()` and `akern.reg()`.

Zeros in the response compositional data are allowed.

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

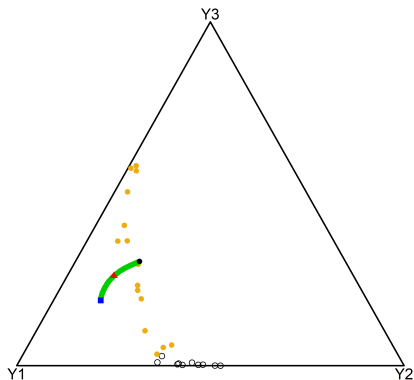


Figure 1: ■ = Fréchet mean with $\alpha = -1$, ▲ = Fréchet mean with $\alpha = 0$ and ● = Fréchet mean with $\alpha = 1$. The dashed green curve - shows the path of all Fréchet means starting with $\alpha = -1$ up to $\alpha = 1$. The golden circles indicate the set of observations used to compute the Fréchet mean.

Simulation Study Set Up

- We compared the predictive performance of the proposed α - k -NN to KLD regression.
- Polynomial and segmented relationships between the response and predictor variables were considered.
- n : Between 100 and 1000 with increasing step size of 50.
- D : 3, 5, 7 and 10.
- Scenarios were repeated with the addition of zero values.
- A 10-fold cv protocol was applied for each regression model to evaluate its predictive performance, measured by the Kullback-Leibler divergence $\sum_{j=1}^n \sum_{i=1}^D y_{ji} \log \frac{y_{ji}}{\hat{y}_{ji}}$.
- Results averaged over 100 repeats.

Simulation study results without zero values present

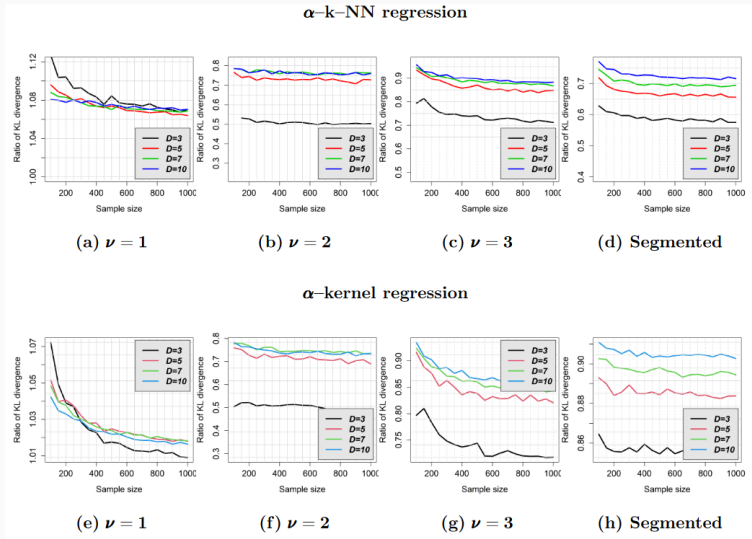


Figure 2: Ratio of KL between the non-parametric regressions and the KLD regression.

Simulation study results with zero values present

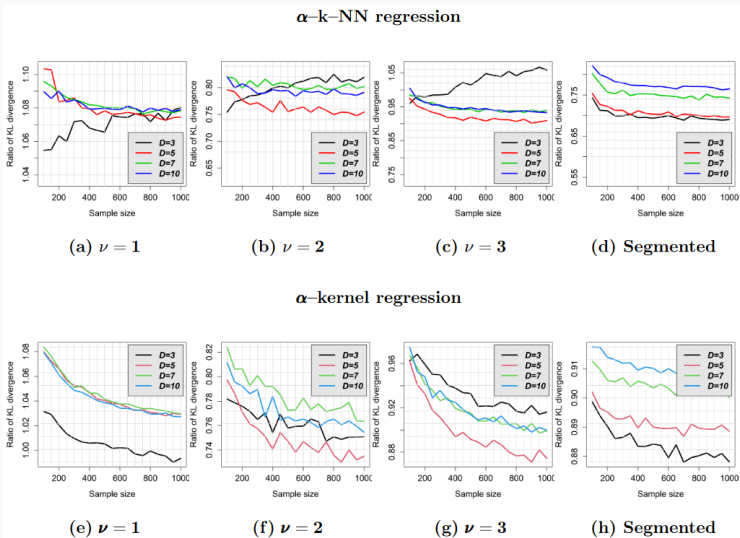


Figure 3: Ratio of KL between the non-parametric regressions and the KLD regression.

Computational cost of the α - k - NN regression

- 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, ..., 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.
- 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).

Computational cost of the α - k - NN regression

Sample size	$D = 7$			$D = 10$		
	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)

Figure 4: Computational cost of OLS, KLD and α - k - NN regression models.

Example with real data

- **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* 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.
- **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* and almost half of the observations (42 out of 92) contain at least one zero value.

Example with real data 1: Gemmas dataset

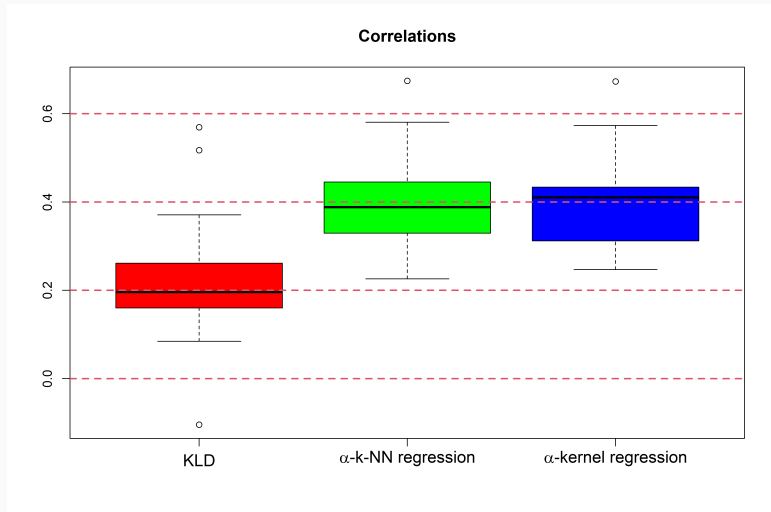


Figure 5: Correlations between the true and the observed components.

Example with real data 2: Glacial dataset

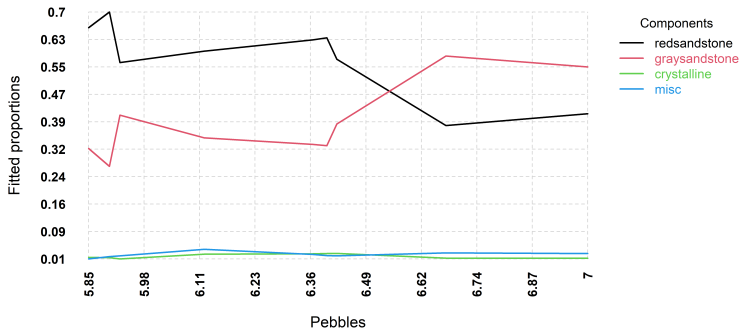







Figure 6: ICE plot of the affect of the predictor on the compositional response when using the α - k -NN regression.

Concluding Remarks

- α - k - NN regression can lead to improvements in accuracy for complex relationships between compositional response data and euclidean predictor variables.
- Compared to KLD regression, α - k - NN regression is much faster for large scale data.
- α -kernel regression, an extension of α - k - NN regression, is an alternative non-parametric approach that provides more flexibility but is less computationally efficient.
- α - k - NN and α -kernel (to a lesser degree) regression lack a framework for classical statistical inference. However, the use of ICE plots can assist in visualizing the predictors' effect on the response compositional data.

-  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.
-  Kendall, W. S. and Le, H. (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.
-  Murteira, J. M. R. and Ramalho, J. J. S. (2016).
Regression analysis of multivariate fractional data.
Econometric Reviews, 35(4):515–552.
-  Tsagris, M., Preston, S., and Wood, A. (2011).
A data-based power transformation for compositional data.
In *Proceedings of the 4rth Compositional Data Analysis Workshop*,