Journal of Data Science,17(1). P. 219 - 238,2019
DOI:10.6339/JDS.201901_17(1).0010

# REGRESSION FOR COMPOSITIONAL DATA WITH COMPOSITIONAL DATA AS PREDICTOR VARIABLES WITH OR WITHOUT ZERO VALUES

Abdulaziz Alenazi*

*Department of Mathematics, Northern Border University, Arar, Saudi Arabia*

## ABSTRACT

Compositional data are positive multivariate data, constrained to lie within the simplex space. Regression analysis of such data has been studied and many regression models have been proposed, but most of them not allowing for zero values. Secondly, the case of compositional data being in the predictor variables side has gained little research interest. Surprisingly enough, the case of both the response and predictor variables being compositional data has not been widely studied. This paper suggests a solution for this last problem. Principal components regression using the $\alpha$-transformation and Kulback-Leibler divergence are the key elements of the proposed approach. An advantage of this approach is that zero values are allowed, in both the response and the predictor variables side. Simulation studies and examples with real data illustrate the performance of our algorithm.

**Keywords:** Compositional data, regression, principal components, α-transformation

* Corresponding author
Email: a.alenazi@nbu.edu.sa

## 1 Introduction

Compositional data are positive multivariate data whose vector elements sum to the same constant usually taken to be 1 for convenience. Data of this type arise in many fields such as geology, ecology, archaeometry, economics, geochemistry, biology, political sciences and forensic sciences, showing their wide applicability. Their support, termed simplex, is given by

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

where $D$ denotes the number of variables (better known as components).

There are various techniques for regression analysis with compositional data being the response variables. See for example Aitchison (2003) who used classical methods on the log -ratio transformed space and Gueorguieva et al. (2008) who applied Dirichlet regression. Stephens (1982) and Scealy and Welsh (2011) transformed the data on the surface of the unit hyper-sphere, using the square root transformation, and thus treat them as directional data. Tsagris (2015b) proposed the $\alpha$-regression which relies upon the $\alpha$-transformation (Tsagris et al., 2011), whereas divergence based regression techniques were suggested by Tsagris (2015a) and Murteira and Ramalho (2016). Finally, compositional data regression from the Bayesian perspective was suggested by Shimizu et al. (2015).

An important issue in compositional data is the presence of zeros, which cause problems for the logarithmic transformation. The issue of zero values in some components is not addressed in most papers and especially in the task of regression. When zero values exist in data, Dirichlet models and the log-ratio transformation suggested by Aitchison (1982, 2003) and Egozcue et al. (2003) will not work unless a zero value imputation is applied first. The square root transformation on the other hand, the $\alpha$-regression and divergence based regression models treat the zero values naturally. More recently, Tsagris and Stewart (2018) proposed a Dirichlet regression modified to account for zero values. As for the classification setting, Tsagris (2014) proposed the use of a power transformation applicable to cases with zero values in the data.

Most papers focus on compositional data being in the response variable side. The case of compositional data in the predictor variables side was treated first by Hron et al. (2012) who applied the isometric log-ratio transformation, defined in (4), to the compositional data and then applied a standard linear regression model. The case of both the dependent and the independent variables containing compositional data has been treated by Wang et al. (2015) in the context of variable selection. Wang et al. (2013) suggested the use of the isometric log -ratio transformation (4) on both sides. Wang et al. (2010) is the one closest to our work who transformed the compositional data using (4) and then applied partial least squares.

The isometric log-ratio transformation (4), or the more general $\alpha$-transformation (6), reduces the dimensionality of the compositional data by 1, via the sub-Helmert matrix (the Helmert matrix (Lancaster, 1965) without the first row). Collinearities in the data may still exist, and this is why Tsagris (2015b) suggested the use of principal components regression (Jolliffe, 2005). A second advantage of the latter approach is that unlike the isometric log -ratio transformation (4), the $\alpha$-transformation (6) is applicable when zero values are present and no zero values imputation is necessary. This is very important, since in large datasets with many zero values, (not necessarily sparse data) imputation is not a suggested strategy. In addition, Tsagris (2015b) showed that $\alpha$-PCR can lead to better predictions, with a value of $\alpha$ other than zero.

We propose a solution combining some of the aforementioned papers. Specifically, we engage the $\alpha$-principal components regression ($\alpha$-PCR) for the independent compositional data and the multinomial logit regression (MLR) model (Murteira and Ramalho, 2016) for the response compositional data. We not only substitute the partial least squares of Wang et al. (2010) with PCA but we also generalize either of these methods since a more general transformation than a logarithmic is applied. In addition, zero values are treated naturally, without any modification of the data or conditional distributions.

In the next section we present some preliminaries regarding the statistical analysis of compositional data, the MLR model (Murteira and Ramalho, 2016) and the $\alpha$-PCR (Tsagris,2015b). The new approach is presented in Section 3. Simulation studies are presented in Section 4 and a demonstration with real data is given in Section 5. Finally, the conclusions close the paper.

# 2 Preliminaries

Below we summarize some important information regarding compositional data that will help us throughout this paper. We start with the listing of some relevant transformations, continuing with the $\alpha$-PCR and the multinomial logit regression.

## 2.1 Transformations for compositional data

### 2.1.1 Additive log-ratio transformation

For a composition $\mathbf{x} \in \mathbb{S}^D$, the additive log-ratio transformation is defined in Aitchison (1982) as

$$y_i = \log\left(\frac{x_i}{x_D}\right) \; for \; i = 1, \dots, D-1, \qquad (2)$$

where $x_D$ is the last component playing the role of the common divisor, the choice of which is not restrictive, since any component can play this role.

### 2.1.2 Centered log-ratio transformation

The drawback of (2) is that it treats the components asymmetrically. The interpretation of the results, in many cases, depends upon the common divisor. For this reason Aitchison (1983) suggested the centered log-ratio transformation

$$w_i = \log\left(\frac{x_i}{\prod_{j=1}^{D} x_j^{1/D}}\right), \text{for } i = 1, \ldots, D. \tag{3}$$

Note that the zero sum constraint in Equation (3) is an obvious drawback of this transformation as it can lead to singularity issues.

### 2.1.3 Isometric log-ratio transformation

In order to remove the redundant dimension imposed by this constraint, one can apply the so called isometric log-ratio transformation (Egozcue et al., 2003)

$$z_0 = \mathbf{wH}^T, \tag{4}$$

where $\mathbf{H}$ is the Helmert matrix (Lancaster, 1965) (an orthonormal $D \times D$ matrix) after deletion of the first row. This is a standard orthogonal matrix in shape analysis used to overcome singularity problems (Dryden & Mardia, 1998;, Le & Small, 1999). By left multiplying the data with the Helmert matrix (without the first row) the data are mapped onto $\mathbb{R}^{D-1}$ and the zero sum constraint is removed.

### 2.1.4 $\boldsymbol{\alpha}$-transformation

Tsagris, Preston & Wood (2011) developed the $\alpha$-transformation as a more general transformation than that in Equation (4). Let

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

denote the power transformation for compositional data as defined by Aitchison (2003). In a manner analogous to Equations (3-4) they defined the $\alpha$-transformation to be

$$\mathbf{z}_\alpha = \left(\frac{D\mathbf{u}_\alpha}{\alpha} - \frac{1}{\alpha}\right)\mathbf{H}^T. \tag{6}$$

The $\alpha$-transformation (6) converges to the isometric log-ratio transformation (4) as $\alpha$ tends to zero (Tsagris et al., 2011).

## 2.2 $\boldsymbol{\alpha}$-Principal components regression

Hron et al. (2012) studied the case of compositional data being predictor variables by using (4). Tsagris (2015b) has already covered this case and we are interested in generalizing this idea to cover the case of multicollinearity as well.

Principal components regression (PCR) is based on principal component analysis (Jolliffe, 2005) and hence we will briefly describe the algorithm for PCR using the $\alpha$-transformation (Tsagris, 2015b):

1. Choose a value of $\alpha$, apply the $\alpha$-transformation (6) onto the compositional data (independent variables) **X** and obtain $\mathbf{Z}_\alpha$.

2. Perform eigen analysis on $\mathbf{Z}_\alpha^T\mathbf{Z}_\alpha$ and calculate the matrix of the eigenvectors **V** and the scores $\mathbf{SC} = \mathbf{Z}_\alpha\mathbf{V}$.

3. Perform regression analysis using the scores (**SC**) as predictor variables.

We will term the above procedure $\alpha$-PCR. The value of $\alpha$ and the number of principal components which lead to the optimal results are chosen via cross validation (CV) described in detail in Section 3.

We can use any number of eigenvectors (or principal component). If we use all of them, then we end up with the fitted values as if we implemented a regression model with all the components of the independent composition. However, our focus is to use a subset of them in order to reduce noise. We do not perform feature selection, nor do we perform any statistical inference regarding the coefficients of the principal components. From the perspective of an applied scientist or a researcher from another field (e.g. ecologist), statistical inference is important. They would be interested in the significance of the predictor variables. But the way we have formulated our approach, the significance of the predictor variables is not feasible. In addition, we are more interested in estimating the dependent compositions as accurately as possible. That is we are more interested in the predictive performance of the model.

## 2.3 Multinomial logit regression

When compositional data are in the response variables side (the usual and most studied case), Murteira and Ramalho (2016) mentioned the use of the Kullback-Leibler divergence of the observed from the fitted compositional vectors

$$\min_{\boldsymbol{\beta}} \sum_{i=1}^{n} y_i \log \frac{y_i}{\mathbf{f_i}(\boldsymbol{\beta}; x)} = \max_{\boldsymbol{\beta}} \sum_{i=1}^{n} y_i \log \mathbf{f_i}(\boldsymbol{\beta}; x), \tag{7}$$

where

$$\mathbf{f_i}(\boldsymbol{\beta}; x) = \left( \frac{1}{\sum_{j=1}^{D} e^{x_i^T \boldsymbol{\beta}_j}}, \frac{e^{x_i^T \boldsymbol{\beta}_2}}{\sum_{j=1}^{D} e^{x_i^T \boldsymbol{\beta}_j}}, \dots, \frac{e^{x_i^T \boldsymbol{\beta}_d}}{\sum_{j=1}^{D} e^{x_i^T \boldsymbol{\beta}_j}} \right)$$

and **y** and **x** are the compositional response variables and the set of predictor variables respectively.

Closed form solution for the minimization of (7) does not exist, but the use of the Newton -Raphson algorithm (Böhning, 1992) can speed up the minimization process. The advantage of using (7) instead of the classical multivariate regression after the additive (2) or the isometric (4) log-ratio transformation, is that zeros can be treated naturally and require no further changes or modifications.

## 3 Compositional-compositional regression

Compositional-compositional regression refers to the case of both the response and the predictor variables consisting of compositional data. We will denote the response or dependent variables by response or dependent composition and the independent variables by independent composition or predictor composition. We have defined all the prerequisites necessary to construct our proposed methodology. All we need to do is couple the $\alpha$-PCR with the MLR.

Hence, we will substitute $\mathbf{x}$ in (7) with $\mathbf{SC} = \mathbf{Z}_{\boldsymbol{\alpha}}\mathbf{V}$, the scores of the PCA after applying the $\alpha$-transformation to the independent composition, and (7) will become

$$\max_{\boldsymbol{\beta}} \sum_{i-1}^{n} y_i \log \mathbf{f}_{\mathbf{i}} \left(\boldsymbol{\beta}; \mathbf{SC}\right),$$

where

$$\mathbf{f}_{\mathbf{i}}(\boldsymbol{\beta}; \mathbf{SC}) = \left( \frac{1}{\sum_{j=1}^{D} e^{\mathrm{SC}_i^T \boldsymbol{\beta}_j}}, \frac{e^{\mathrm{SC}_i^T \boldsymbol{\beta}_2}}{\sum_{j=1}^{D} e^{\mathrm{SC}_i^T \boldsymbol{\beta}_j}}, \ldots, \frac{e^{\mathrm{SC}_i^T \boldsymbol{\beta}_d}}{\sum_{j=1}^{D} e^{\mathrm{SC}_i^T \boldsymbol{\beta}_j}} \right).$$

Similarly to Tsagris (2015b) we will use the *K*-fold CV protocol to choose the optimal values of $\alpha$ and $s$, the number of principal components. According to the *K*-fold CV protocol we split the data into *K* distinct sets, and each time remove one fold and use *K*−1 sets to build the compositional-compositional regression. The fold remained outside the model construction is used as a test set to validate the model. The performance or predictability of the model is measured by the Kullback-Leibler divergence of the true from the fitted compositional vectors. The optimal results in our case, chosen values of $\alpha$ and $s$, the number of Principal Components (PCs), will refer to minimization of the mean Kullback-Leibler divergence of the observed from the predicted compositional data.

The response and the predictor compositions need not be of the same dimensions and the use of the MLR is by no means restrictive. One could also use the usual multivariate linear regression model (Mardia et al., 1979) on the log-ratio transformed data, Dirichlet regression (Gueorguieva et al., 2008) and its adjusted version for zero values (Tsagris and Stewart, 2018), the ordinary least squares (Murteira and Ramalho, 2016), the ES-OV regression (Tsagris, 2015a) or any other regression model for compositional response variables. We propose the MLR model though because not only it handles zero values naturally, but also because it is very fast to fit.

## 4 Simulation studies

We conducted simulation studies to assess the performance of the proposed algorithm. We examined our algorithm's capability of recovering the true values of $\alpha$ and the true number $s$. We used the open software R (R Core Team, 2017) for all our simulation studies an data analysis.

We used the *fgl* dataset available in the *MASS* library in R (Ripley, 2002). The data concern measurements of forensic glass fragments and consist of 214 observations on 8 chemical elements. The dataset contains a huge amount of zeros and suits excellent for our purpose. We chose a value of $\alpha$ and applied the $\alpha$-transformation (6). In the transformed data, we calculated the 7 PCs. We took the scores from the first PC and multiplied it with some generated beta coefficients. We then generated data from a multivariate normal and added some white noise to them. The data were then mapped onto the simplex using the inverse of the additive log-ratio transformation (2). Keeping the value of $\alpha$ constant we repeated this procedure using all 7 PCs, one by one. We then chose another value of $\alpha$ and repeated this process.

In all cases, positive values of $\alpha$ were used only and specifically we used 10 values, ranging from 0.1 up to 1 equally spaced. The above process was repeated 200 times for each combination of $\alpha$ and number of PCs. For each combination we computed the difference between the true and the estimated value of $\alpha$, and the percentage the times the algorithm chose the correct value of $s$. Tables 1 and 2 summarize our findings and Figure 1 illustrates them.

Table 1 contains the average bias of $\alpha$ when the dependent composition was generated from one, two or three PCs. The estimated bias of $\alpha$ is always small, and is not really affected by the true value of $\alpha$. When 4 or more PCs were used, the bias increases.

Table 2 contains the proportion of correct identification of the number of PCs. Whether our methodology selected the true number of PCs used does not depend so much on the true value of $\alpha$, but rather on the number of PCs. By examining the table in column-wise manner, we see that as we move from left to right, the percentage of correct identification of the true number of PCs decreases.

Table 1: Estimated bias $\sum_{i=1}^{200}(\hat{\alpha}_i - \alpha)/200$ of the 10-fold CV for a range of values of $\alpha$ and number of PCs.

| | PCs=1 | PCs=2 | PCs=3 | PCs=4 | PCs=5 | PCs=6 | PCs=7 |
|---|---|---|---|---|---|---|---|
| $\alpha = 0.1$ | 0.001 | 0.027 | 0.049 | 0.076 | 0.084 | 0.093 | 0.072 |
| $\alpha = 0.2$ | 0.001 | 0.026 | 0.033 | 0.017 | 0.03 | 0.061 | 0.079 |
| $\alpha = 0.3$ | -0.001 | 0.002 | 0.01 | 0.016 | -0.004 | 0.008 | -0.014 |
| $\alpha = 0.4$ | 0.002 | 0.006 | -0.024 | 0.041 | -0.078 | -0.021 | -0.047 |
| $\alpha = 0.5$ | 0.006 | 0.005 | 0.018 | 0.017 | -0.079 | -0.098 | -0.070 |
| $\alpha = 0.6$ | 0.001 | -0.002 | 0.021 | -0.021 | -0.07 | -0.091 | -0.128 |
| $\alpha = 0.7$ | -0.001 | -0.028 | 0.004 | -0.070 | -0.09 | -0.127 | -0.131 |
| $\alpha = 0.8$ | -0.001 | -0.012 | -0.043 | -0.023 | -0.128 | -0.083 | 0.012 |
| $\alpha = 0.9$ | 0.002 | 0.001 | -0.005 | -0.003 | -0.055 | -0.027 | 0.008 |
| $\alpha = 1$ | 0.001 | 0.001 | -0.02 | -0.057 | -0.072 | -0.087 | -0.097 |

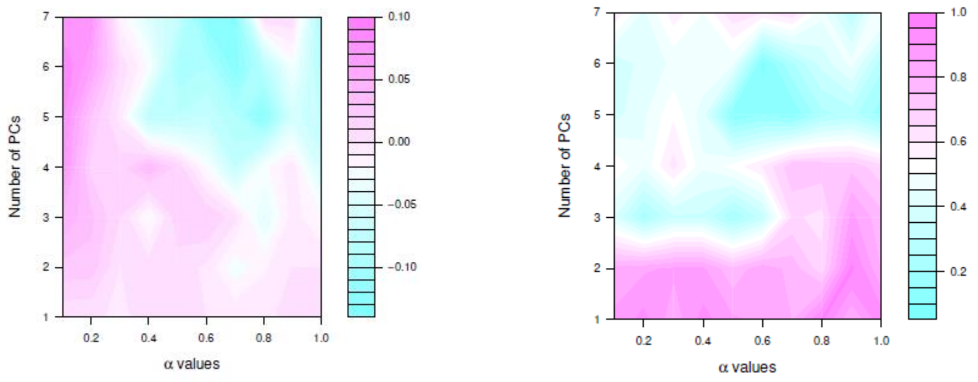
Table 2: Proportion of times the correct number of PCs was selected by the 10-fold CV for a range of values of $\alpha$ and number of PCs.

| | PCs=1 | PCs=2 | PCs=3 | PCs=4 | PCs=5 | PCs=6 | PCs=7 |
|---|---|---|---|---|---|---|---|
| $\alpha = 0.1$ | 0.86 | 0.82 | 0.4 | 0.54 | 0.38 | 0.36 | 0.60 |
| $\alpha = 0.2$ | 0.92 | 0.84 | 0.22 | 0.46 | 0.44 | 0.42 | 0.44 |
| $\alpha = 0.3$ | 0.82 | 0.86 | 0.34 | 0.62 | 0.54 | 0.46 | 0.58 |
| $\alpha = 0.4$ | 0.92 | 0.86 | 0.34 | 0.46 | 0.42 | 0.5 | 0.50 |
| $\alpha = 0.5$ | 0.86 | 0.76 | 0.2 | 0.52 | 0.1 | 0.46 | 0.64 |
| $\alpha = 0.6$ | 0.84 | 0.84 | 0.32 | 0.6 | 0.14 | 0.08 | 0.60 |
| $\alpha = 0.7$ | 0.8 | 0.78 | 0.66 | 0.74 | 0.1 | 0.24 | 0.64 |
| $\alpha = 0.8$ | 0.96 | 0.67 | 0.63 | 0.73 | 0.23 | 0.37 | 0.47 |
| $\alpha = 0.9$ | 0.8 | 0.97 | 0.87 | 0.73 | 0.27 | 0.47 | 0.30 |
| $\alpha = 1$ | 0.93 | 0.8 | 0.77 | 0.67 | 0.17 | 0.27 | 0.53 |

# 5 Examples with real data

## 5.1 Example datasets

We will now introduce four datasets, taken from Aitchison (2003), to illustrate the performance of our proposed approach described in Section 3. The sample sizes (see Table 3) are rather small. Hence, we will not perform a 10-fold CV, but a leave-one-out CV



(a) (b)

Figure 1: Graphical representation (heatmap plots) of Tables 1 and 2. (a) Estimated bias $\sum_{i=1}^{200}(\hat{\alpha}_i - \alpha)\,/200$ for a range of values of $s$. (b) Proportion of correct choice number of PCs.

(LOOCV) protocol and we will report the optimal pairs of parameters (values of $\alpha$ and $s$) found.

- **Clam ecology**. From the many colonies of clams in East Bay, 20 colonies from the East Bay were selected at random and from each a sample of clams was taken. For each colony the proportions of clams in each colour-size combination was estimated and the corresponding compositions, consisting of 6 components, were recorded. A similar study was conducted in West Bay and the resulting 20 colour-size compositions. The task of interest is to quantify the relationship between the two compositions.
- **Hair and eye colours**. For each of the 33 counties of Scotland the percentages of boys in five hair-colour categories and four eye-colour categories are available. The question of interest is to predict the composition of the hair colours from the eye colour compositions and vice versa.
- **White-cells**. A cytologist is interested in the possibility of introducing into his laboratory a new method of determining the white-cell composition of a blood sample, that is the proportions of the three kinds of white cells, among the total of white cells observed. The current method involves time-consuming, microscopic inspection and is known to be accurate, whereas the proposed method is a quick automatic image analysis whose accuracy is still largely undetermined. In an experiment to assess the effectiveness of the proposed method, each of 30 blood samples was halved, one half being assigned randomly to one method, the other half to the other method. It is fairly obvious that the two methods produce different compositions and we are interested in predicting the microscopic inspection compositions from the compositions produced by the new method.
- **Fruit evaluation**. The yatquat tree produces each season a single large fruit whose quality is assessed in terms of the relative proportions by volume of flesh, skin and stone. In an experiment to investigate whether a certain hormone influences quality, an agricultural

scientist uses 40 yatquat trees, randomly allocates 20 trees to the hormone treatment and leaves untreated the remaining 20 trees. The fruit compositions of the 40 trees in the present and the preceding season are available. The task is to examine the relationship between the compositions in the two seasons.

In the *Clam ecology* and *Hair and eye colours* datasets there is no dependent and independent variables. Hence, we will perform regression on both ways. In the *White-cells* dataset, the response composition is the microscopic inspection and the independent composition is the new method of image analysis. For the *Fruit evaluation dataset*, the current season is the response composition and the past season will play the role of the predictor composition.

Table 3: Information about the example pairs of datasets used. The sample size and the number of components for each composition is given.

| Dataset | Sample size | No of components<br>of the one composition | No of components<br>of the other composition |
|---|---|---|---|
| Clam ecology | 20 | 6 | 6 |
| Hair and eye colour | 33 | 5 | 4 |
| White-cells | 30 | 3 | 3 |
| Fruit evaluation | 40 | 3 | 3 |

### 5.2 Results

Figure 2 contains the heatmap plots of LOOCV using the first two datasets (*Clam ecology and Hair and eye colours*), whereas Figure 3 contains the heatmap plots of LOOCV using the other two datasets (*White-cells and Fruit evaluation*). Table 4 contains the optimal values of $\alpha$ and $s$, along with their corresponding minimum Kullback-Leibler divergence. We see that in most cases a value of $\alpha$ other than zero is the optimal for transforming the predictor compositions.

For the clam ecology dataset, only one PC is necessary to reach the optimal predictive performance for both cases. The chosen value of $\alpha$ is not the same whatsoever, whereas for the *hair and eye colour* dataset the combination of $\alpha$ and $s$ is similar in both cases. In these two datasets, there was no distinction between response and predictor composition, but for the *white cells* and *fruit evaluation* dataset there is. For the *white-cells*, the $\alpha = 0.2$ and $s = 2$ PCs lead to the optimal predictive performance, while for the *fruit evaluation* the optimal transformation was the $\alpha$-transformation with $\alpha = 1$ using $s = 1$ PC. This means, that the data were not $\alpha$-transformed, but centered and multiplied by a constant, prior to the Helmert multiplication.

Let us now make an observation for the *Clam-ecology II* example dataset, where the optimal value of $\alpha$ was $-1$. We remind the reader, that if zero values are present in the independent composition, only positive values of $\alpha$ are allowed.

Table 4: Results for each dataset. The optimal value of $\alpha$ and number of PCs along with the minimum average Kullback-Leibler divergence.

| Dataset | $\alpha$ | Number of PCS | Average Kullback-Leibler divergence |
|---|---|---|---|
| Clam ecology I | 0.0 | 1 | 0.022 |
| Clam ecology II | -1 | 1 | 0.023 |
| Hair and eye colour I | 0.5 | 2 | 0.004 |
| Hair and eye colour II | 0.4 | 3 | 0.005 |
| White-cells | 0.2 | 2 | 0.005 |
| Fruit evaluation | 1.0 | 1 | 0.0015 |

## 6 Conclusions

In this work we suggested a methodology for the case of both the response and the predictor variables consisting of compositional data. The approach is based upon combining prior work for regression with compositional data. Our simulation studies and examples illustrated the proposed methodology showing that it works satisfactorily. We should highlight also, that instead of PCA or partial least squares (Wang et al., 2010) other methods could be used as well. In any case, the $\alpha$-transformation (6) should be applied to the independent composition as it gives more flexibility than the isometric log-ratio transformation (4) and allows for zero values.

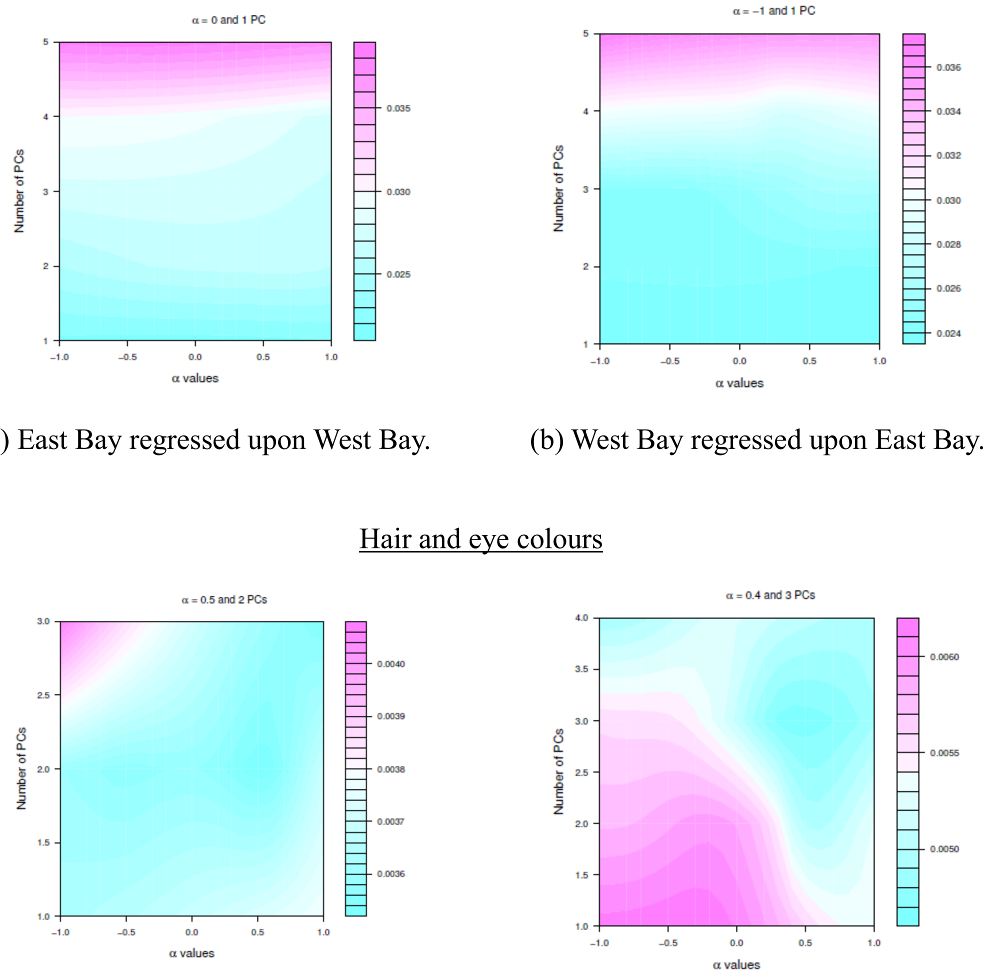
The multinomial logit regression is not the only available regression model. In the case of no zero values, the usual log-ratio methodology or any other regression model could be used as well. In addition, principal components is again not the only dimensionality reduction technique. Principal coordinate analysis, or kernel PCA (Schölkopf et al., 1997), more general, could be used to capture non-linear dependencies among the compositional data.

Our simulation studies indicated that the bias in the estimated $\alpha$ increases as the true number of PCs required increases. A similar pattern was observed when selecting the number of PCs. The higher the true number of PCs, the lower the probability of selecting their exact number.

The examples with real data analysis illustrated the performance of our proposed methodology.

Our future research is oriented at exploring the performance of our proposed method-ology in higher dimensions, for either the response or the predictor composition. This

Clam ecology



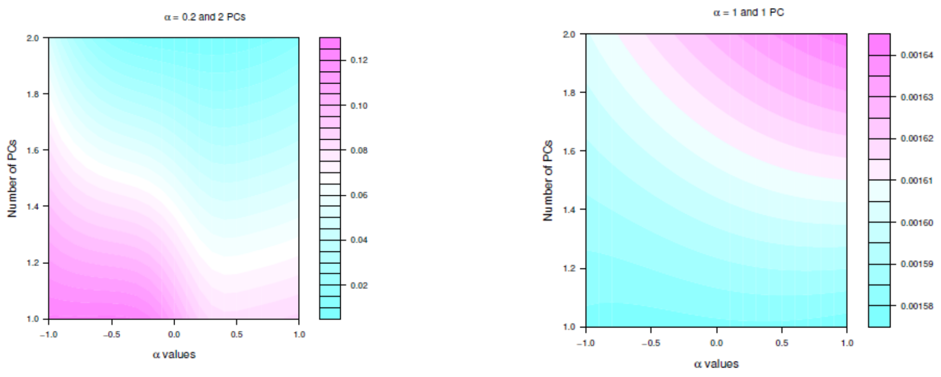
(a) East Bay regressed upon West Bay. (b) West Bay regressed upon East Bay.

Hair and eye colours

(a) Hair colour regressed upon eye colour. (b) Eye colour regressed upon hair colour.

Figure 2: Heatmap plots of the Clam ecology and Hair and eye colours datasets. The horizontal axis contains the $\alpha$ values and the vertical axis the number of PCs used. The values of the Kullback-Leibler divergence are plotted, with low values being desirable.

issue has been examined only recently by Li (2015); Fang et al. (2015) and Kaul et al. (2016). In addition, the robustness to outliers and the case of zero values is to be investigated as well. Outliers in the predictor composition can be addressed via robust PCA



(a) White-cells.

(b) Fruit evaluation.

Figure 3: Heatmap plots of the White-cells and Fruit evaluation datasets. The horizontal axis contains the $\alpha$ values and the vertical axis the number of PCs used. The values of the Kullback-Leibler divergence are plotted, with low values being desirable.

(Filzmoser et al., 2018), while outliers in the response composition can be addressed via substitution of the MLR with a robust multivariate regression model after applying the $\alpha$-transformation (6).

## Appendix

### $\alpha$-transformation and MLR

The package *Compositional* is required.

```
kl.alfapcr <- function(y, x, covar = NULL, a, k, xnew = NULL,
  B = 1, ncores = 1, tol = 1e-07, maxiters = 50) {
  z <- Compositional::alfa(x, a)$aff
  n <- nrow(z)
  p <- ncol(z)
  if (k > p) {
    k <- p
}
eig <- Compositional::prcomp(z, center = FALSE, scale = FALSE)
values <- eig$sdev^2
per <- cumsum( values / sum(values) )
vec <- eig$rotation[, 1:k, drop=FALSE]
sc <- eig$x[, 1:k, drop = FALSE]
if ( !is.null(covar) ) {
    sc <- cbind(sc, covar)
}
if ( !is.null(xnew) ) {
    xnew <- Compositional::alfa(xnew, a)$aff
    xnew <- cbind(xnew %*% vec, covar)
}
Compositional::kl.compreg(y, sc, xnew = xnew, B = B, ncores = ncores,
       tol = tol, maxiters = maxiters)
}
```

### Tuning of the $\alpha$-transformation and the number of PCs used in the MLR

```
klalfapcr.tune <- function(y, x, covar = NULL, M = 10, maxk = 50,
a = seq(-1, 1, by = 0.1), mat = NULL, graph = FALSE,
tol = 1e-07, maxiters = 50) {
    n <- nrow(x)
    p <- nncol(x) – 1
    if ( min(x) = = 0 ) {
      a <- a [ a > 0 ]
```

```
}
if ( maxk > p ) {
  maxk  <- p
}
if ( !is.null(covar) )    {
covar  <- as.matrix(covar)
}
if ( is.null(mat) ) {
nu  <- sample(1:n, min( n, round(n / M) * M ) )
options(warn = -1)
mat  <- matrix( nu, ncol = M )
} else {
 mat  <- mat
}
M  <- ncol(mat)
rmat  <- nrow(mat)
mspe  <- list()
msp  <- matrix( nrow = M, ncol = maxk )
colnames(msp)  <- paste("PC", 1:maxk, sep = " ")
  for ( i in 1:length(a) ) {
  xa  <- Compositional::alfa(x, a[i])$aff
  for (vim in 1:M) {
    ytest  <- y[ mat[, vim], , drop = FALSE ]
    ytrain  <- y[ -mat[, vim], , drop = FALSE ]
    xtrain  <- xa[ -mat[, vim],,       drop = FALSE ]
    xtest  <- xa[ mat[, vim], , drop = FALSE ]
    com  <- sum(ytest * log(ytest), na.rm = TRUE)
    mod  <- Compositional::prcomp(xtrain, center = FALSE)
    vec  <- mod$rotation
    za  <- mod$x
    zanew  <- xtest %*% vec
    for ( j in 1:maxk ) {
      if ( !is.null(covar) ) {
       z  <- cbind(za[, 1:j, drop = FALSE],
       covar[ -mat[, vim], drop = FALSE ] )
       znew  <- cbind(zanew[, 1:j, drop = FALSE],
                  covar[ mat[, vim], drop = FALSE ] )
```

```
          } else {
            z  <- za[, 1:j, drop = FALSE ]
              znew  <- zanew[, 1:j, drop = FALSE]
          }
          est  <- Compositional::kl.compreg(y = ytrain, x = z, xnew = znew,
    tol = 1e-07, maxiters = 50)$est
          res  <- sum(ytest * log(est), na.rm = TRUE)
          msp[vim, j]  <- com - res * is.finite(res)
        }
      }
      mspe[[ i ]]  <- msp
    }
names(mspe)  <- paste("alpha=", a, sep = "")
performance  <- lapply(mspe, colMeans)
performance  <- matrix( unlist(performance), ncol = maxk, byrow = TRUE )
colnames(performance) <- paste("PC", 1:maxk, sep = " ")
rownames(performance) <- paste("alpha", a, sep = " ")
poia  <- which(performance = = min(performance, na.rm = TRUE), arr.ind =TRUE) params
<- c( a[ poia[, 1] ], poia[, 2] )
names(params) <- c("best alpha", "best k")
if ( graph ) {
  filled.contour(a, 1:maxk, performance,
                 xlab = expression(paste(alpha, " values")),
                 ylab = "Number of PCs", cex.lab = 1.3 )
}
list(mspe = mspe, performance = performance,
     best.perf = min(performance), params = params)
}
```

## References

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

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

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

[4] B ö hning, D. (1992). Multinomial logistic regression algorithm. A*nnals of the Institute of Statistical Mathematics*, 44(1):197–200.

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

[6] Fang, H., Huang, C., Zhao, H., and Deng, M. (2015). Cclasso: correlation inference for compositional data through lasso. *Bioinformatics*, 31(19):3172–3180.

[7] Filzmoser, P., Fritz, H., and Kalcher, K. (2018). *pcaPP*: Robust *PCA* by *Projection Pursuit [Software]*.

[8] Gueorguieva, R., Rosenheck, R., and Zelterman, D. (2008). Dirichlet component regression and its applications to psychiatric data. *Computational statistics & data analysis*, 52:5344–5355.

[9] Hron, K., Filzmoser, P., and Thompson, K. (2012). Linear regression with compositional explanatory variables. Journal of *Applied Statistics*, 39(5):1115–1128.

[10] Jolliffe, I. T. (2005). Principal *component analysis*. New York: Springer-Verlag.

[11] Kaul, A., Davidov, O., and Peddada, S. D. (2016). Analysis of high dimensional compositional data containing structural zeros with applications to microbiome data. *arXiv preprint arXiv*:1605.06193.

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

[13] Li, H. (2015). Microbiome, metagenomics, and high-dimensional compositional data analysis. *Annual Review of Statistics and Its Application*, 2:73–94.

[14] Mardia, K., Kent, J., and Bibby, J. (1979). *Multivariate Analysis*. London: Academic Press.

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

[16] R Core Team (2017). R: A *Language and Environment for Statistical Computing, version* 3.4. R Foundation for Statistical Computing, Vienna, Austria.

[17] Ripley, B. (2002). Modern applied statistics with s.*Statistics and Computing,fourth ed.Springer, New York*.

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

[19] Schölkopf, B., Smola, A., and Mu¨ller, K.-R. (1997). Kernel principal component analysis. *In International Conference on Artificial Neural Networks*, pages 583–588. Springer.

[20] Shimizu, T. K., Louzada, F., Suzuki, A. K., and Ehlers, R. S. (2015). Modeling Compositional Regression with uncorrelated and correlated errors: a Bayesian approach. *Journal of Data Science*, 16(2):221–250.

[21] Stephens, M. A. (1982). Use of the von Mises distribution to analyse continuous proportions. *Biometrika*, 69(1):197–203.

[22] Tsagris, M. (2014). The k-nn algorithm for compositional data: a revised approach with and without zero values present. *Journal of Data Science*, 12(3):519–534.

[23] Tsagris, M. (2015a). A novel, divergence based, regression for compositional data. In *Proceedings of the 28th Panhellenic Statistics Conference, Athens, Greece, pp. 430–44*.

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

[25] Tsagris,M.and Stewart, C. (2018). A Dirichlet Regression Model for Compositional Data with Zeros. *Lobachevskii Journal of Mathematics*, 39(3):398 –412.

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

[27] Wang, H., Huang, L., Shangguan, L., and Wang, S. (2015). Variable selection and estimation for regression models with compositional data predictors. In *6th International Workshop on Compositional Data Analysis, Girona, Spain*.

[28] Wang, H., Meng, J., and Tenenhaus, M. (2010). *Regression Modelling Analysis on Compositional Data*, pages 381–406. Springer Berlin Heidelberg.

[29] Wang, H., Shangguan, L., Wu, J., and Guan, R. (2013).Multiple linear regression modeling for compositional data. *Neurocomputing*, 122:490–500.