Constrained least squares simplicial-simplicial regression

Michail Tsagris

Department of Economics, University of Crete,
Gallos Campus, Rethymnon, Greece
mtsagris@uoc.gr

March 29, 2024

Abstract

Simplicial-simplicial regression refers to the regression setting where both the responses and predictor variables lie within the simplex space, i.e. they are compositional. For this setting, constrained least squares, where the regression coefficients themselves lie within the simplex, is proposed. The model is transformation-free but the adoption of a power transformation is straightforward, it can treat more than one compositional datasets as predictors and offers the possibility of weights among the simplicial predictors. Among the model's advantages are its ability to treat zeros in a natural way and a highly computationally efficient algorithm to estimate its coefficients. Resampling based hypothesis testing procedures are employed regarding inference, such as linear independence, and equality of the regression coefficients to some pre-specified values. The performance of the proposed technique and its comparison to an existing methodology that is of the same spirit takes place using simulation studies and real data examples.

Keywords: compositional data, regression, quadratic programming

1 Introduction

Compositional data[1] are non-negative multivariate vectors whose variables (typically called components) conveying only relative information. When the vectors are scaled to sum to 1, their sample space is the standard simplex

$$\mathbb{S}^{D-1} = \left\{ (y_1, ..., y_D)^\top \middle| y_i \geq 0, \sum_{i=1}^D y_i = 1 \right\}, \tag{1}$$

where D denotes the number of components.

Examples of such 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[2].

The widespread occurrence of this type of data in numerous scientific fields that involve predictors has necessitated the need for valid regression models which in turn has led to several developments in this area, many of which have been proposed recently. Most of these regression models have a restricted

[1] In the field of econometrics they are known as multivariate fractional data (Mullahy, 2015, Murteira and Ramalho, 2016).

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

attention to the case of a simplicial response (simplicial-real regression setting), or a simplicial predictor (real-simplicial regression setting). The case of simplicial-simplicial regression, where both sides of the equation contain compositional data has not gained too much attention, and this is the main focus of this paper.

Most published papers regarding the last case scenario involve transformations of both simplicial sides. Hron et al. (2012), Wang et al. (2013), Chen et al. (2017) and Han and Yu (2022) used a log-ratio transformation for both the response and predictor variables and performed a multivariate linear regression model. Alenazi (2019) transformed the simplicial predictor using the α-transformation (Tsagris et al., 2011) followed by principal component analysis and then employed the Kullback-Leibler divergence regression (or multinomial logit) model (Murteira and Ramalho, 2016). The exception is Fiksel et al. (2022) who proposed a transformation-free linear regression (TFLR) model whose coefficients lie within the simplex and are estimated via minimization of the Kullback-Leibler divergence (KLD) between the observed and the fitted simplicial responses.

An important issue with compositional data analysis is the presence of zeros that prohibit the use of the logarithmic transformations, and hence the approach of Hron et al. (2012), Wang et al. (2013) and Chen et al. (2017), an issue that is not addressed in most papers. The classical strategy addressing this issue is to replace the zero values by a small quantity (Aitchison, 2003). However, the approach of Alenazi (2019) and handles the zero cases in a natural manner. This is not true in general for the TFLR model though.

Tsagris et al. (2011) categorized the compositional data analysis approaches into two main categories, the raw data approach and the log-ratio approach. A perhaps better classification would be the raw data and the transformation-based approaches. Moving along the lines of the raw data approach the paper proposes the use of the same transformation-free linear regression model, as in Fiksel et al. (2022), when both the response and the predictor variables are simplicial. However, the adoption of a power transformation in the simplicial response generalizes the model. The regression coefficients are estimated via simplicial constrained least squares (SCLS) and as the name implies, least squares is the loss function used to estimate the regression coefficients which are constrained to lie on the simplex. This in turn implies that the expected value of the simplicial response can be expressed as a Markov transition from the simplicial predictor. The proposed SCLS model allows for more than one simplicial predictor, further allows the possibility of assigning weights to the simplicial predictors, and treats zero values naturally, in both the simplicial response and the predictor variables. The assumption of linear independence between the simplicial variables, and hypotheses regarding the mateix of regression coefficients can be tested using resampling techniques. Evidently, the SCLS is similar in spirit to the TFLR, but they have different loss (or objective functions). The TFLR model employs the Expectation-Maximization (EM) algorithm, whereas the SCLS model is based on quadratic programming, thus it enjoys a really low computational cost.

The problem of constrained least squares (CLS), with a univariate real response, is not new. Liew (1976) and Wets (1991) have studied the asymptotic properties of constrained regression and have established the consistency of the regression coefficients, assuming the linear specification is correct. Wets (1991) specifically formalized the asymtptotic properties of the regression coefficients for the case of the M-estimators, whose least squares is a special case. More recently, James et al. (2019) proposed the constrained LASSO, a penalized version of the CLS. The current work though differs from these works in that it deals with the case of a constrained multivariate response.

The rest of the paper is structured as follows. Section 2 reviews some simplicial-simplicial regression models, while Section 3 introduces the SCLS model, and discusses several cases related to the TFLR model as well. Section 4 contains Monte-Carlo simulation studies comparing the SCLS to the TFLR model in terms of a) type I and type II errors of the linear independence assumption, b) discrepancy of the estimated regression coefficients, and c) computational cost. The SCLS model is then applied to real data for illustration, and comparison to the TFLR model, purposes, while the last section concludes the paper.

2 Review of simplicial-simplicial regression models

Two commonly used log-ratio transformations, as well as a more general α–transformation are defined, followed by some existing regression models for simplicial-simplicial regression.

2.1 Log-ratio simplicial-simplicial regression models

Aitchison (1982) suggested applying the additive log-ratio (alr) transformation to compositional data prior to using standard multivariate data analysis techniques. Let $\mathbf{y} = (y_1, \ldots, y_D)^\top \in \mathbb{S}^{D-1}$, then the alr transformation is given by

$$\mathbf{v} = \{v_j\}_{j=1,\ldots,D-1} = \left\{\log \frac{y_j}{y_1}\right\}_{j=2,\ldots,D}, \tag{2}$$

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.

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

$$\mathbf{u} = \{u_j\}_{j=1,\ldots,D} = \left\{\log \frac{y_j}{g\left(\mathbf{y}\right)}\right\}_{j=1,\ldots,D,} \tag{3}$$

where $g\left(\mathbf{y}\right) = \prod_{j=1}^{D} y_j^{1/D}$ is the geometric mean.

The clr transformation (3) was proposed in the context of principal component analysis with the potential drawback that $\sum_{j=1}^{D} u_j = 0$, so essentially the unity sum constraint is replaced by the zero sum constraint. To overcome the singularity problem, Egozcue et al. (2003) proposed multiplying Eq. (3) 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

$$\mathbf{z}_0 = \mathbf{H}\mathbf{u}, \tag{4}$$

where $\mathbf{z}_0 = \left(z_{0,1}, \ldots, z_{0,D-1}\right)^\top \in \mathbb{R}^{D-1}$. Note that $\mathbf{H}$ may be replaced by any orthogonal matrix which preserves distances (Tsagris et al., 2011).

Simplicial-simplicial regression based on the ilr tranformation (Chen et al., 2017, Han and Yu, 2022, Hron et al., 2012, Wang et al., 2013) is similar to alr regression and is carried out by transforming both the response and the predictor variables via the alr (2) or the ilr (4) transformations

$$\mathrm{E}(\mathbf{v}_k|\mathbf{X}) = \beta_{0k} + \sum_{j=1}^{D_p-1} \beta_{jk}\mathrm{alr}(\mathbf{X}_j) \ \text{ and } \ \mathrm{E}(\mathbf{z}_{0k}|\mathbf{X}) = \beta_{0k} + \sum_{j=1}^{D_p-1} \beta_{jk}\mathrm{ilr}(\mathbf{X}_j), \quad k = 1, \ldots, D_r - 1,$$

where D_r and D_p denote the number of components of the response and compositional variables, respectively.

Moving along the same lines, (Wang et al., 2009) proposed partial least squares regression where both the response and the predictor variables are first transformed using the ilr transformation (4), and the PLS is applied to the transformed data. Kernel regression (Di Marzio et al., 2015), such as the Nadaraya-Watson or local polynomial regression, may also be applied in a similar manner.

The fitted values for both the alr and ilr transformations are the same and are therefore generally back transformed onto the simplex using the appropriate inverse transformation for ease of interpretation. The drawback of these regression models is their inability to handle zero values in the compositional response data. The popular solution is to apply zero substitution strategies (Aitchison, 2003, Martín-Fernández et al., 2012) prior to fitting these regression models.

2.2 Kullback-Leibler divergence principal component regression

Zero imputation strategies strategy, however, can produce regression models with predictive performance worse than regression models that can treat zeros naturally (Tsagris, 2015a). When zeros occur in the data or more flexibility is required, the Box-Cox type transformation proposed by Tsagris et al. (2011) may be employed. Specifically, Aitchison (2003) defined the power transformation as

$$\mathbf{w}_\alpha = \left\{ \frac{y_j^\alpha}{\sum_{l=1}^{D} y_l^\alpha} \right\}_{j=1,\ldots,D} \tag{5}$$

and Tsagris et al. (2011) subsequently defined the α-transformation, based on (5), as

$$\mathbf{z}_\alpha = \frac{1}{\alpha}\mathbf{H}\left(D\mathbf{w}_\alpha - \mathbf{1}_D\right), \tag{6}$$

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 Eq. (5) remains in the simplex $\mathbb{S}^{D-1}$, $\mathbf{z}_\alpha$ in Eq. (6) is mapped onto a subset of $\mathbb{R}^{D-1}$. Furthermore, as $\alpha \to 0$, Eq. (6) converges to the ilr transformation[3] in Eq. (4) (Tsagris et al., 2016), provided no zero values exist in the data. 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 benefit of the α-transformation over the alr and clr transformations is that it can be applied even when zero values are present in the data (using strictly positive values of α), offer more flexibility and yield better results (Tsagris, 2015b, Tsagris et al., 2016, Tsagris and Stewart, 2022).

In the context of simplicial-real regression, Murteira and Ramalho (2016) minimize the KLD

$$\min_{\boldsymbol{\beta}} \sum_{i=1}^{n} \mathbf{y}_i^\top \log \frac{\mathbf{y}_i}{\hat{\mathbf{y}}_i} = \max_{\boldsymbol{\beta}} \sum_{i=1}^{n} \mathbf{y}_i^\top \log \hat{\mathbf{y}}_i, \tag{7}$$

where n denotes the sample size of the $\mathbf{u}_i$ the observed simplicial response data and $\hat{\mathbf{u}}_i = (\hat{y}_{i1}, \ldots, \hat{y}_{iD})^\top$ are the fitted simplicial response data which have been transformed to simplex space through the transformation

$$\hat{y}_{ij} = \left\{ \begin{array}{cc} \frac{1}{1+\sum_{l=2}^{D_r} e^{\mathbf{x}_i^\top \boldsymbol{\beta}_l}} & \text{if} \quad j = 1 \\ \frac{e^{\mathbf{x}_i^\top \boldsymbol{\beta}_j}}{1+\sum_{l=2}^{D_r} e^{\mathbf{x}_i^\top \boldsymbol{\beta}_l}} & \text{for} \quad j = 2, ..., D_r, \end{array} \right\} \tag{8}$$

where where $\mathbf{x}_i^\top$ denotes the i-th row of the design matrix $\mathbf{X}$ containing the, non-simplicial, predictor variables, and $\boldsymbol{\beta}_j = (\beta_{0j}, \beta_{1j}, ..., \beta_{pj})^\top$, $j = 2, \ldots, D$ (Tsagris, 2015a,b).

The KLD regression model in Eq. (7), also referred to as Multinomial logit regression is a semi-parametric regression technique, and unlike alr and ilr regression it can handle zeros naturally, since $\lim_{x\to 0} x \log x = 0$.

Alenazi (2019) performed principal component analysis (Jolliffe, 2005) on the α-transformed simplicial predictor $\mathbf{X}$, and then used the projections onto the first k principal components as Euclidean predictors to the KLD regression model. This approach focuses on improving the predictive performance of the alr and ilr regression models, while increasing the computational cost, as the value of α and the number k of the principal components must be tuned via cross-validation[4].

2.3 The TFLR model

The TFLR model (Fiksel et al., 2022) relates the simplicial response to the simplicial predictor via a linear transformation

$$E(\mathbf{Y}_k|\mathbf{X}) = \sum_{j=1}^{D_p} B_{jk}\mathbf{X}_j, \tag{9}$$

[3]The scaling factor D exists to assist in the convergence.

[4]For a comparison of this approach to the TFLR model, the reader is addressed to Fiksel et al. (2022).

where $\mathbf{B}$ itself belongs to the simplex, $\mathbf{B} \in \mathbb{R}^{D_p \times D_r} | B_{jk} \geq 0, \sum_{k=1}^{D_r} B_{jk} = 1$. Estimation of the elements of $\mathbf{B}$ takes place by minimizing the KLD as in Eq. (7)

$$\min \left\{ \sum_{i=1}^{n} \sum_{k=1}^{Dr} y_{ik} \log \left(\frac{y_{ik}}{\sum_{j=1}^{D_p} B_{jk} x_{ij}} \right) \right\} = \max \left\{ \sum_{i=1}^{n} \sum_{k=1}^{Dr} y_{ik} \log \left(\sum_{j=1}^{D_p} B_{jk} x_{ij} \right) \right\}. \tag{10}$$

The aforementioned approach allows for zero values in the response variable, but not for the simplicial predictor, in general. Think for example the case of an observation $\mathbf{x}_i = (x_{i1}, x_{i2}, 0, 0)$, where x_{i1} and x_{i2} are different from zero and and the k-th column of the estimated matrix $\mathbf{B}$ is equal to $(0, 0, B_{3k}, B_{4k})^\top$, where both B_{3k} and B_{4k} are not zero. Another possibility is when a column of the $\mathbf{B}$ matrix is full of zeros. These two cases will produce $x_{ik} B_{ik} = 0$ and hence (10) cannot be computed[5].

3 The simplicial constrained least squares model

The SCLS model adopts the same link as TFLR (9) between the simplicial response and predictor variables, but only this time the elements of $\mathbf{B}$ are estimated by minimizing the squared loss

$$\begin{aligned} \text{SL}(\mathbf{B}) &= \sum_{i=1}^{n} \sum_{k=1}^{D_r} \left(y_{ik} - \sum_{j=1}^{D_p} B_{jk} \mathbf{X}_j \right)^2 = \text{tr} \left(\mathbf{Y} - \mathbf{XB} \right) \left(\mathbf{Y} - \mathbf{XB} \right)^\top \\ &\propto 2 \left[-\text{tr} \left(\mathbf{Y}^\top \mathbf{XB} \right) + \frac{1}{2} \text{tr} \left(\mathbf{B}^\top \mathbf{X}^\top \mathbf{XB} \right) \right]. \end{aligned} \tag{11}$$

3.1 The SCLS model and quadratic programming

Quadratic programming solves the following problem

$$\min_{\boldsymbol{b}} \left\{ -\boldsymbol{d}^\top \boldsymbol{b} + \frac{1}{2} \boldsymbol{b}^\top \mathbf{D} \boldsymbol{b} \right\}, \quad \text{under the constraints} \quad \mathbf{A}^\top \boldsymbol{b} \geq \boldsymbol{b}_0. \tag{12}$$

The SL($\mathbf{B}$) though (11) is minimized via quadratic matrix programming, but it can be formulated under the quadratic programming framework[6], as in the constrained minimization formula in Eq. (12). The $\mathbf{D}$ matrix is a $D_{rp} \times D_{rp}$ diagonal matrix, where $D_{rp} = D_r \times D_p$, and is related to the $\mathbf{X}^\top \mathbf{X}$ in the following manner

$$\mathbf{D} = \mathbf{I}_{D_r} \otimes \mathbf{X}^\top \mathbf{X} = \begin{pmatrix} \mathbf{X}^\top \mathbf{X} & \mathbf{0} & \dots & \mathbf{0} \\ \mathbf{0} & \ddots & \mathbf{0} & \vdots \\ \vdots & \mathbf{0} & \ddots & \mathbf{0} \\ \mathbf{0} & \dots & \mathbf{0} & \mathbf{X}^\top \mathbf{X} \end{pmatrix}, \tag{13}$$

where $\otimes$ denotes the Kronecker product and $\mathbf{I}_{D_r}$ is the $D_r \times D_r$ identity matrix.

The $\mathbf{A}^\top$ matrix can be broken down two three sub-matrices, the $D_r \times D_{rp}$ $\mathbf{A}_1$, the $D_{rp} \times D_{rp}$ $\mathbf{A}_2$ and the the $D_{rp} \times D_{rp}$ $\mathbf{A}_3$, one stacked under the other,

$$\mathbf{A}^\top = \begin{pmatrix} \mathbf{A}_1^\top \\ \mathbf{A}_2^\top \\ \mathbf{A}_3^\top \end{pmatrix} = \mathbf{A}^\top = \begin{pmatrix} \mathbf{I}_{D_r} & \dots & \mathbf{I}_{D_r} \\ & \mathbf{I}_{D_{rp}} & \\ & -\mathbf{I}_{D_{rp}} & \end{pmatrix}.$$

[5]This phenomenon may occur with the KLD regression model as well but it is rather highly unlikely.

[6]The SCLS model along with all relevant functions used throught this paper has been implemented in the *R* package *Compositional* (Tsagris et al., 2024) which makes use of the *R* package *quadprog* (Turlach et al., 2019) that has implemented the algorithm of Goldfarb and Idnani (1983) to minimize Eq. (11).

- The $\mathbf{A}_1^\top$ matrix contains the identity matrix $\mathbf{I}_{D_r}$, D_p times, one next to the other

$$\mathbf{A}_1^\top = \left(\mathbf{I}_{D_r}, \ldots, \mathbf{I}_{D_r}\right).$$

  This is related to the unity sum constraint of the row coefficients. This means that

$$\mathbf{A}_1^\top \mathbf{b} = \left(\mathbf{I}_{D_r}, \ldots, \mathbf{I}_{D_r}\right)\mathbf{b} = \left(1, \ldots, 1\right)^\top.$$

- The $\mathbf{A}_2^\top$ matrix is the identity matrix $\mathbf{I}_{D_rp}$

$$\mathbf{A}_2^\top = \mathbf{I}_{D_{rp}}.$$

  This is related to the fact that all coefficients take non-negative values, thus

$$\mathbf{A}_2^\top \mathbf{b} = \mathbf{I}_{D_{rp}} \mathbf{b} \geq \left(0, \ldots, 0\right)^\top.$$

- The $\mathbf{A}_3^\top$ matrix is the negated identity matrix $-\mathbf{I}_{D_rp}$

$$\mathbf{A}_3^\top = -\mathbf{I}_{D_{rp}}.$$

This is related to the fact that the values of all coefficients must be less than or equal to 1, and hence

$$\mathbf{A}_3^\top \mathbf{b} = -\mathbf{I}_{D_{rp}} \mathbf{b} \geq \left(-1, \ldots, -1\right)^\top \quad \left(\text{or } \; \mathbf{I}_{D_{rp}} \mathbf{b} \leq \left(1, \ldots, 1\right)^\top\right).$$

The $\boldsymbol{b}$ relates to the $\mathbf{B}$ matrix via the vectorization operation $\boldsymbol{b} = \text{vec}\left(\mathbf{B}\right)$, that is, each column of the matrix is stacked one under the other. Finally, the $\boldsymbol{b}_0$ vector contains the D_p-dimensional vector of 1s, $\boldsymbol{j}_{D_p}$, that corresponds to the unity sum constraint of the rows of $\mathbf{B}$ (linked to the $\mathbf{A}_1^\top$ matrix), the D_{rp}-dimensional vector of 0s, $\mathbf{0}_{D_{rp}}$, corresponding to the fact that all elements B_{jk} of $\mathbf{B}$ are non-negative (linked to the $\mathbf{A}_2^\top$ matrix) and finally the D_{rp}-dimensional vector of -1s, $-\boldsymbol{j}_{D_{rp}}$, corresponding to the fact that all elements B_{jk} of $\mathbf{B}$ are at bounded by unity from above (linked to the $\mathbf{A}_3^\top$ matrix)

$$\boldsymbol{b}_0 = \left(\boldsymbol{j}_{D_p}, \mathbf{0}_{D_{rp}}, -\boldsymbol{j}_{D_{rp}}\right)^\top.$$

3.2 Interpretation of the regression coefficients

Interpretation of the resulting estimated coefficients is a crucial aspect of a regression model if one is interested in making inference about them[7]. Tsagris et al. (2023) for instance proposed some non-parametric regression models that do not estimate coefficients, and thus visualized the (non-linear) effects of the predictor variables graphically. SCLS on the other hand yields regression coefficients and being a transformation free regression model, interpretation of its regression coefficients has similarities to the interpretation of the coefficients of the regression model with the alr (2) transformation. In the latter, a change in a predictor variable refers to a relative change in the components of the simplicial response, and in the SCLS, the interpretation is somewhat similar. If x_j increases by δ, while x_l decreases by the same amount, ceteris paribus, the expected change of $\mathbf{y}$ is equal to $\delta\left(\mathbf{B}_j - \mathbf{B}_l\right)$, where j and l denote rows. Assume for instance the following form of the matrix of regression coefficients

$$\mathbf{B} = \begin{pmatrix} 0.20 & 0.40 & 0.40 \\ 0.10 & 0.30 & 0.60 \\ 0.30 & 0.35 & 0.35 \\ 0.50 & 0.40 & 0.30 \end{pmatrix} \tag{14}$$

If the first component of the predictor increases by 0.1, while the second component decreases by the same amount, the expected change in the response is equal to $0.1\left(0.2 - 0.1, 0.4 - 0.3, 0.4 - 0.6\right) = \left(0.01, 0.01, -0.02\right)$. While this interpretation is easy to understand it entails a perhaps minor downside. The interpretation is not universally applicable, as an expected change may lead to a point outside the simplex.

[7]Hypothesis testing for the parameters relies upon resampling techniques (presented next) due to the lack of parametric assumptions imposed on the coefficients.

3.3 Visualization of the regression coefficients

In the case of a 3-part simplicial response, the coefficients can also be visualized using the ternary diagram, as in Figure 1, that contains the 4 rows of the **B** matrix in Eq. (14). The interpretation of the ternary diagram is as follows. A point close to a vertex indicates high proportion in that vertex-associated component, while a point close to an edge opposite a vertex, indicates a low value in the vertex-associated component and finally a point that lies close to the barycentre of the triangle indicates almost equal values in all components. The third value of the second row of the matrix (B_{23}) is equal to 0.60 and thus closer to the top vertex, whereas the third row values are similar, and hence B_3 lies close to the barycentre of the triangle. When the (row-)coefficients of a component of the simplicial predictor take values to the barycenter, i.e. $(1/D_r, \ldots, 1/D_r)$ this indicates that the particular component of the simplicial predictor does not carry too much statistical information. On the contrary, the further the coefficients associated with a component of the simplicial predictor are from the barycentre, the more information this component carries. In this simple example, the third component does not seem very important, because the third row coefficients (B_3) are close to the barycentre. The second component (B_2) of the simplicial predictor on the other hand, is probably the most important component, because its vector of coefficients lie the furthest from the barycentre among the other three vectors of coefficients.

3.3.1 Confidence intervals for the regression coefficients

Confidence intervals for the true parameters can be constructed using non-parametric bootstrap (Fiksel et al., 2022) but the drawback is that the lower and upper values do not sum to unity. Confidence regions on the other hand are more intuitive. They are relatively easy to compute for 3-part simplicial responses and can be visualized in a ternary diagram in that case[8]. Another non-parametric option would be to use empirical likelihood (Owen, 2001), but to produce simultaneous confidence regions for all sets of regression coefficients would be computationally hard, especially as the dimensionality of the simplicial predictor, D_p, increases. To ease the computational burden one could produce profile confidence regions instead. Following Fiksel et al. (2022), the minimum volume ellipsoid (Van Aelst and Rousseeuw, 2009) that contains the 95% of the bootstrap based coefficients is estimated and its ellipsoid hull is computed using the algorithm of Pison et al. (1999).

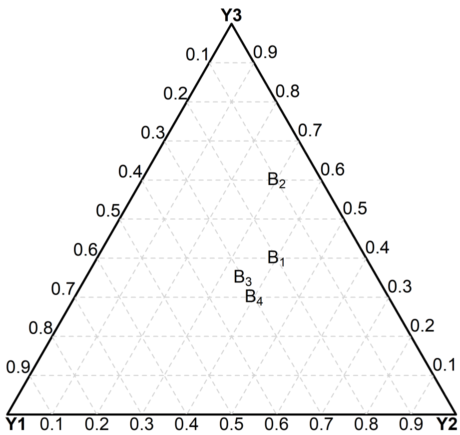


Figure 1: Ternary plot of the example matrix of coefficients **B** of Eq. (1).

[8]In case of 4 components one could use a trinagular pyramid that has four equilateral triangles with all edges equal in length.

3.4 A test of linear independence

Following Fiksel et al. (2022) a test of linear independence between $\mathbf{Y}$ and $\mathbf{X}$ that relies upon permutations is proposed. If $\mathbf{Y}$ is linearly independent of $\mathbf{X}$, then $\mathrm{E}(\mathbf{Y}|\mathbf{X}) = \mathrm{E}(\mathbf{Y})$. The test statistic utilized by Fiksel et al. (2022) is based upon the KLD, but in the SCLS model it will be based upon the $\mathrm{SL}^r(\mathbf{B})$. The steps to compute the permutation based p-value are delineated below.

Step 1 Compute $\mathrm{SL}^{obs}(\mathbf{B})$ using the observed data $\mathbf{Y}$ and $\mathbf{X}$.

Step 2 Create $\mathbf{X}^r$ by permuting at random the rows of $\mathbf{X}$.

Step 3 Compute $\mathrm{SL}^r(\mathbf{B})$ using $\mathbf{Y}$ and $\mathbf{X}^r$.

Step 4 Repeat Steps 2 and 3 R times (e.g. $R = 999$).

Step 5 Calculate the permutation based p-value $= \frac{\sum_{r=1}^R I\left[\mathrm{SL}^r(\mathbf{B}) \leq \mathrm{SL}^{obs}(\mathbf{B})\right]+1}{R+1}$.

3.5 A test for specific values imposed on regression coefficients

In order to test the null hypothesis that $\mathbf{B} = \mathbf{B}_0$ versus the alternative that at least one of the elements of $\mathbf{B}$ differs, the permutation technique may be employed as well. One such example is the case of $B_{j1} = \ldots = B_{jD_r} = 1/D_r$ for some j, $1 \leq j \leq D_p$, indicating that the j-th component is not important. Below are two examples where this test could be applied.

Amalagamation of components refers to summing the values of at least two components. For instance take the simplicial vector $\mathbf{x} = (x_1, x_2, x_3, x_4)$. An example of an amalgamation would be $\mathbf{x}^* = (x_1, x_2 + x_3, x_4)$. In the regression case this would imply that for some given components of the simplicial predictor, e.g. $1 \leq l_1, l_2 \leq D_p$ and $l_1 \neq l_2$, the expected value of the simplicial response is written as

$$E(\mathbf{Y}_k|\mathbf{X}) = \sum_{j \neq l_1, l_2}^{D_p} B_{jk}\mathbf{X}_j + B_{l_1k}\mathbf{X}_{l_1} + B_{l_2k}\mathbf{X}_{l_2} = \sum_{j \neq l_1, l_2}^{D_p} B_{jk}\mathbf{X}_j + B_{l_1k}\left(\mathbf{X}_{l_1} + \mathbf{X}_{l_2}\right), \tag{15}$$

This implies to test for equality of two rows of the matrix $\mathbf{B}$, for instance that $\mathbf{B}_{l_1} = \mathbf{B}_{l_2}$, where $\mathbf{B}_l$ denotes the l-th row of the matrix $\mathbf{B}$. The steps are similar to the linear test of independence where the difference is noticed at Step 2. Instead of permuting the rows of the simplicial predictor $\mathbf{X}$, its two columns, the components $(\mathbf{x}_{l_1}, \mathbf{x}_{l_2})$ are permuted. The rows of $\mathbf{X}$ are kept constant, but those two columns are permuted at random.

If this hypothesis is not rejected it implies that these two components can be amalgamated and form a new simplicial predictor with $D_p - 1$ components. Evidently, the same procedure could be applied to the simplicial response and in that case one ends up with an amalgamated simplicial response and in that case the hypothesis would refer to the equality of two columns of the matrix $\mathbf{B}$.

A sub-composition stems from a composition by removing at least one component and rescaling the rest of the values to some to unity. The resulting null hypothesis in this regression setting is $B_{j1} = \ldots = B_{jD_r} = 0$ for some j, $1 \leq j \leq D_p$. However, there is an associated problem with this, the rescaling part, which in order to be valid the property of sub-compositional coherence[9] (Aitchison, 2003) should be met and in this case it is not. However, the theoretical implications of the sub-compositionality are not known, and hence this test perhaps requires more investigation.

The hypothesis $B_{1k} = \ldots = B_{D_pk} = 0$ for some k, $1 \leq k \leq D_r$ is also meaningful as it implies that the k-th component of the simplicial response is not affected by any component of the simplicial predictor. Note however that this hypothesis cannot be tested by the TFLR for the reason of introducing zeros to the fitted values.

[9]This property refers to the invariance of the results when the full composition or the sub-composition is used.

3.6 More than one simplicial predictors

The SCLS model can be extended to the case of M ($M > 1$) simplicial predictors, in which case Eq. (9) may be written as

$$E(\mathbf{Y}_k|\mathbf{X}^1, \ldots, \mathbf{X}^M) = \sum_{m=1}^{M} \sum_{j=1}^{D_p} \frac{B_{jk}^m}{M} \mathbf{X}_j^m, \tag{16}$$

where the division of the matrix of coefficients $\mathbf{B}$ by M ensures that the estimated simplicial responses sum to unity.

Fiksel et al. (2022) did not examine this case, but TFLR can be applied to this case as well. The drawback of the SCLS model in this case is that the $\mathbf{D}$ matrix (13) will not be positive definite, a drawback which can be handled by the algorithm of Higham (2002) (using Dykstra's correction) that forces positive definiteness. The algorithm yields an approximate solution which is deemed satisfactory.

The implication of Eq. (16) is that each simplicial predictor carries equal weight. One can escape this restrictive assumption by assigning weights to the regression coefficient matrix of each set of simplicial predictors, and hence write Eq. (16) as

$$E(\mathbf{Y}_k|\mathbf{X}^1, \ldots, \mathbf{X}^M) = \sum_{m=1}^{M} \sum_{j=1}^{D_p} a_m B_{jk}^m \mathbf{X}_j^m, \tag{17}$$

where $a_m \geq 0$ for $m = 1, \ldots, M$ and $\sum_{m=1}^{M} = 1$, are the weights assigned to each set of simplicial predictors. Thus, the previous case (16) can be seen as a special case of (17) where $a_1 = \ldots = a_m = 1/M$. The weights may be computed using quadratic programming again, but in doing so the dimensionality of the required matrix $\mathbf{D}$ will explode as the number of simplicial predictors and the number of components of each increase, thus one can use simpler optimizers in R. The possibility of allowing for more than one simplicial predictors in either model opens the door to conditional association testing and subsequently to simplicial variable selection.

3.7 The α-SCLS and α-TFLR models

Another possible extension is to apply the power transformation (5) to the simplicial response.

$$E(\mathbf{W}_\alpha|\mathbf{X}) = \mathbf{X}\mathbf{B}, \tag{18}$$

and their predictions are then back-transformed using the inverse of (5)

$$\hat{\mathbf{y}} = \left\{ \frac{\hat{w}_j^{1/\alpha}}{\sum_{l=1}^{D} \hat{w}_l^{1/\alpha}} \right\}_{j=1,\ldots,D_r}.$$

The extension will be denoted α-SCLS and α-TFLR models and evidently when $\alpha = 1$ the SCLS and TFLR models, respectively, are recovered. The power transformation strategy was shown to improve the accuracy in the classification setting (Tsagris, 2014), at the cost of introducing the interpretation-predictive performance trade-off. The cost of hard to interpret estimated regression parameters is compensated by the benefit of an increased predictive performance.

3.8 Examples of application of the SCLS model

Similarly to the TFLR, the SCLS (and subesequently the α-SCLS and α-TFLR) can also be applied to a series of other regression settings.

3.8.1 AR(1) model

In the AR(1) model formulation, the current expected simplicial response is a linear function of its lagged time response, $E(\mathbf{Y}_t|\mathbf{Y}_{t-1}) = \mathbf{Y}_{t-1}\mathbf{B}$, thus the matrix $\mathbf{B}$ can be interpreted as the matrix of transition probabilities of the states of the simplicial response at time $t-1$. The linear independence test in this case translates to testing the assumption of no auto-correlation. The case of allowing for more than one simplicial predictors extends the 1 lag to p lags, thus extend the AR(1) model to the AR(p) model. However, there is no theory to this end, what are implications of such a model (i.e. unit root testing, etc.) and thus this requires more research.

3.8.2 Categorical predictor or response

The case of a categorical predictor can be addressed via the SCLS. In that case, each row of the simplicial predictor is written as $\mathbf{x}_i = \mathbf{e}_j$, where $\mathbf{e}_j$ denotes a vector with 0 D_p elements except its j-th element that is 1[10], yielding a one-way analysis of variance type of model.

The linear independence test translates to to testing the assumption of equal simplicial population means. The latter case was studied by Tsagris et al. (2017) using parametric and non-parametric procedures. However, simulation studies (not presented) replicating a case scenario from Tsagris et al. (2017) comparing the SCLS and TFLR with the bootstrap version of the two sample James test showed that the latter is size correct, but the former two produced inflated type I errors. Hence, in contrast to the suggestion made by Fiksel et al. (2022), there is evidence against the use of SCLS and TFLR models for the purpose of hypothesis testing of equality of simplicial means.

Fiksel et al. (2022) stated that when the response is categorical this is equivalent to performing multinomial linear regression, with an identity link. The task is also equivalent to discriminant analysis for which better, non-linear, alternatives exist (Lu et al., 2024, Tsagris et al., 2016).

4 Simulation studies

Simulation studies were conducted to compare the SCLS and TFLR regression models with the axes of comparison being a) the type I and II errors, b) the discrecpancy of the estimated matrix of coefficients, and c) the computational cost (running time). For the first two axes, the simplicial response consisted of $D_r = (3, 5, 7, 10)$ components, while the simplicial predictor contained only 3 components, and the sample sizes considered were $n = (50, 100, 200, 300, 500)$, but for the third axis, the sample size was larger. The results were averaged over $1,000$ replicates for each combination of dimensionality and sample size.

4.1 Type I error

Conforming to the null hypothesis of linear independence, both the values of the simplicial response ($\mathbf{Y}$) and predictor ($\mathbf{X}$) were generated from Dirichlet distributions, independently of one another. The values of the simplicial response were generated from a $\text{Dir}(\mathbf{a})$, whose parameters $\mathbf{a}$ were drawn from the uniform distribution $U(1, 5)$, while the values of the simplicial predictor were generated from the $\text{Dir}(1, 1, 1)$. The estimated type I error for both models is presented in Table 1, where evidently the results are similar, the permutation test attains the correct size, regardless of the model utilized, sample size and the dimensionality of the simplicial response.

[10] This could also be seen as an extreme case of compositional data, where one component takes a value of 1 and all other component have zero values.

Table 1: Estimated type I error of both models, under the null hypothesis of linear independence.

		Number of response components			
Sample size	Model	$D_r = 3$	$D_r = 5$	$D_r = 7$	$D_r = 10$
n=50	SCLS	0.046	0.050	0.055	0.045
	TFLR	0.045	0.054	0.049	0.039
n=100	SCLS	0.038	0.058	0.050	0.040
	TFLR	0.038	0.059	0.054	0.040
n=200	SCLS	0.058	0.047	0.057	0.050
	TFLR	0.053	0.056	0.062	0.056
n=300	SCLS	0.046	0.043	0.050	0.044
	TFLR	0.046	0.036	0.051	0.047
n=500	SCLS	0.046	0.046	0.052	0.051
	TFLR	0.042	0.043	0.045	0.042

4.2 Type II error

The response now was linked to the predictor in a linear manner, as in Eq. (9) with some pre-determined values for $\mathbf{B}$ depending on D_r. Random vectors $\mathbf{x}_i$, for $i = 1, \ldots, n$, were generated from $\text{Dir}\,(1, 1, 1)$, then transformed into $\boldsymbol{\mu}_i = \mathbf{x}_i\mathbf{B}$ and finally random vectors $\mathbf{y}_i$ were generated from $\text{Dir}\,(5\mu_1, \ldots, 5\mu_{D_r})$[11].

The estimated powers for both models are presented in Table 2. For the small sample sized data, TFLR seems to produce higher estimated power levels than the SCLS, but as soon as the sample size increases their estimated power is equal.

Table 2: Estimated power of both models.

		Number of response components			
Sample size	Model	$D_r = 3$	$D_r = 5$	$D_r = 7$	$D_r = 10$
n=50	SCLS	0.977	1.000	0.996	0.942
	TFLR	0.995	1.000	1.000	1.000
n=100	SCLS	1.000	1.000	1.000	1.000
	TFLR	1.000	1.000	1.000	1.000
n=200	SCLS	1.000	1.000	1.000	1.000
	TFLR	1.000	1.000	1.000	1.000
n=300	SCLS	1.000	1.000	1.000	1.000
	TFLR	1.000	1.000	1.000	1.000
n=500	SCLS	1.000	1.000	1.000	1.000
	TFLR	1.000	1.000	1.000	1.000

4.3 Discrepancy of the estimated matrix of coefficients

The discrepancy of the estimated matrix of coefficients was assessed by the KLD of the estimated to the true values of the $\mathbf{B}$ matrix, and the L_1 distance between the estimated and the true values of the $\mathbf{B}$

[11]For more information on the specific values of $\mathbf{B}$, the reader is referred to the Appendix.

matrix

$$\text{KLD}(\hat{\mathbf{B}}, \mathbf{B}) = \sum_{j=1}^{D_p}\sum_{k=1}^{D_r} \hat{B}_{jk} \log\left(\frac{\hat{B}_{jk}}{B_{jk}}\right) \quad \text{and} \quad L_1(\hat{\mathbf{B}}, \mathbf{B}) = \sum_{j=1}^{D_p}\sum_{k=1}^{D_r} |\hat{B}_{jk} - B_{jk}|.$$

Table 3 contains the estimated discrepancy quantities. Both discrepancy metrics exhibit small differences between the two models, but the TFLR produces better results. However, the differences in the discrepancies are rather small and can be asserted that the two models are not substantially different. Secondly, for a given dimensionality of the simplicial response, as the sample size increases the differences, using both measures, between the two competing models diminish. Lastly, the KLD increases as the dimensionality of the simplicial responses increases, for a given sample size, however, the L_1 metric decreases, exhibiting a rather unexpected behaviour.

Table 3: Estimated discrepancy of the regression coefficients of the SCLS and TFLR models.

		$\text{KLD}(\hat{\mathbf{B}}, \mathbf{B})$				$L_1(\hat{\mathbf{B}}, \mathbf{B})$			
	Model	$D_r = 3$	$D_r = 5$	$D_r = 7$	$D_r = 10$	$D_r = 3$	$D_r = 5$	$D_r = 7$	$D_r = 10$
n=50	SCLS	0.0096	0.0208	0.0166	0.0163	0.0533	0.0379	0.0414	0.0307
	RFLR	0.0087	0.0145	0.0143	0.0130	0.0514	0.0348	0.0400	0.0290
n=100	SCLS	0.0048	0.0122	0.0088	0.0087	0.0380	0.0271	0.0299	0.0218
	TFLR	0.0044	0.0085	0.0077	0.0070	0.0365	0.0249	0.0287	0.0206
n=200	SCLS	0.0024	0.0075	0.0046	0.0050	0.0274	0.0201	0.0211	0.0158
	TFLR	0.0022	0.0051	0.0040	0.0039	0.0263	0.0181	0.0202	0.0148
n=300	SCLS	0.0014	0.0054	0.0031	0.0036	0.0215	0.0166	0.0172	0.0131
	TFLR	0.0013	0.0037	0.0027	0.0029	0.0207	0.0149	0.0164	0.0123
n=500	SCLS	0.0008	0.0037	0.0021	0.0024	0.0170	0.0132	0.0136	0.0103
	TFLR	0.0007	0.0025	0.0018	0.0019	0.0162	0.0118	0.0129	0.0095

4.4 Computational cost

To evaluate the computational cost the command *benchmark()*, from the *R* package *Rfast2* (Papadakis et al., 2023), was used, based on 20 repetitions. Simplicial response values with $D_r = (3, 5, 7, 10)$ and simplicial predictor values with $D_p = 3$ were generated from a Dirichlet distribution and the sample sizes considered were $n = (500, 1000, 5000, 10000, 20000)$. The speedup factors of the SCLS compared to TFLR are presented in Table 4. The ratios clearly depict that the time required to fit the TFLR model is significantly higher compared to that of the SCLS model.

The TFLR model has been implemented in the *R* package *codalm* (Fiksel and Datta, 2021) and the code is written in *R*, wheras the SCLS model uses the *R* package *quadprog* (Turlach et al., 2019) that relies upon Fortran. So, the running time comparison is not completely fair. Secondly, the code in the TFLR implementation is not highly optimized. To address the second issue, the comparison took place using a self implementation of the TFLR model. In this implementation, the estimated regression coefficients from the SCLS model were used as starting values in the EM algorithm. The new speed-up factors of the SCLS compared to TFLR also appear in Table 4. For the low dimensionality of the simplicial response, the self implementation can be up to 10 times faster than the *codalm*'s implementation, whereas for a higher dimensionality, the improvement drops to only 4 times.

A *C++* implementation is expected to make the TFLR algorithm even faster. However, the number of computations involved in the EM algorithm is still high compared to the quadratic programming approach. Apart from the computational burden, the TFLR model may break down for some combination of zeros in the simplicial predictor and the matrix of coefficients **B** as mentioned in Section 2.3, whereas the SCLS model treats those cases naturally.

Table 4: Speedup factors of SCLS compared to TFLR: ratio of running time of the TFLR model to the running time of the SCLS model.

	Number of response components							
	codalm implementation				Self implementation			
Sample size	$D_r = 3$	$D_r = 5$	$D_r = 7$	$D_r = 10$	$D_r = 3$	$D_r = 5$	$D_r = 7$	$D_r = 10$
n=500	52.430	41.701	58.552	68.708	11.216	15.406	20.166	28.943
n=1000	69.645	85.490	67.583	94.403	12.156	21.820	27.049	66.684
n=5000	121.677	114.959	160.351	261.529	12.151	51.776	37.861	78.321
n=10000	201.123	222.886	159.677	231.467	24.744	37.547	59.314	83.515
n=20000	196.560	329.418	207.237	213.228	26.963	41.129	42.529	56.916

5 Real data analysis

Two real data sets are used to illustrate the performance of the SCLS and compare it with that of TFLR. The first data set comes from the field of agricultural economics while the second comes from the field of political sciences. A third data set was then used to illustrate the confidence regions (of the SCLS model) of the row coefficients of the matrix **B**. Lastly, two of these datasets were further utilized to illustrate the performance of the α-SCLS.

5.1 Crop cultivated area and crop production

Data regarding crop productivity in the Greek NUTS II region of Thessaly during the 2017-218 cropping year were supplied by the Greek Ministry of Agriculture, also known as farm accountancy data network (FADN) data. The data refer to a sample of 487 farms and initially they consisted of 20 crops, but after aggregation they were narrowed down to 10 crops[12]. For each of the 487 farms the cultivated area and the production in each of the 10 crops is known. However, the goal of the paper is to relate the composition of the production (simplicial response, **Y**) to the composition of the cultivated area (simplicial predictor, **X**) and for this reason were scaled to sum to unity[13]. Table 5 contains information about the data: the component names and their simple arithmetic averages.

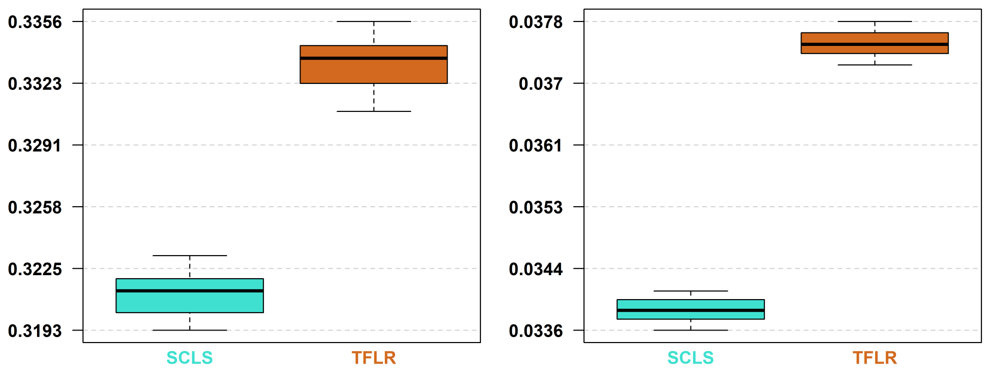
Tables 6 contains the regression coefficients estimated via the SCLS and TFLR models. Evidently, the diagonal coefficients take high values as expected, as the proportion of each crop's cultivated area is expected to be highly related to the proportion of the same crop production. As expected, the linear independence hypothesis is rejected, based on both models. The assumption of a diagonal matrix of coefficients $\mathbf{B} = \mathbf{I}_{10}$ was also rejected by both models.

In order to assess the fit of the two models the KLD of the observed to the predicted simplicial response values and the Jensen-Shannon divergence (JSD) were employed. The KLD was 0.0267 for both models, whereas the JSD values were 0.025 and 0.028 for the SCLS and the TFLR models, respectively. To estimate the predictive performance of each model the 10-fold cross-validation (CV) was utilised and the performance metrics were again the KLD and the JSD. The 10-fold CV was repeated 20 times and the box-plots appearing in Figure 2 visualize the results. The average KLD values were 0.321 and 0.33 for the SCLS and the TFLR models, respectively, whereas the average JSD values were 0.034 and 0.038 for the SCLS and the TFLR models, respectively.

[12]A larger version of this dataset was used in (Mattas et al., 2024).
[13]The raw data cannot be distributed due to disclosure restrictions.

Table 5: Information on the FADN dataset.

Component	Cultivated area (**X**)	Production (**Y**)
X1: Other Cereals	0.1063	0.0958
X2: Durum Wheat	0.1539	0.1407
X3: Maize	0.0600	0.1015
X4: Potatoes, Protein Crops and Rice	0.0393	0.0230
X5: Cotton	0.2315	0.2087
X6: Tobacco, Oil Seeds, Industrial Crops and Vegetables	0.0449	0.0708
X7: Green Plants, Pasture and Grazing	0.1777	0.1777
X8: Fruits, Berries and Nuts	0.0945	0.1098
X9: Olive Trees	0.0733	0.0446
X10: Grapes and Wine	0.0186	0.0273



(a) KLD (b) JSD

Figure 2: FADN data: box-plots of the predictive KLD and JSD for the SCLS and TFLR models.

5.2 Catalan elections

The data set contains the votes in Catalan elections from year 1980 up to 2006 for 41 regions each year. The main parties consist of 6 candidates, while there are votes for other candidates, blank votes and null votes. In total there are 328 observations with 9 variables, that were scaled to sum to 1. The goal here is to assume an AR(1) model, where the lag refers to the one time period between two consecutive election years. Table 7 contains the averages in the 9 components classified per year of election.

Since the task of interest is to perform a time series analysis, the AR(1) SCLS and TFLR models were fitted to the data from the years 1980 up to 2003, and the data from the year of elections 2006 was considered to be the test set. Table 8 contains the estimated coefficients of the SCLS and TFLR models. The prediction capabilities showed that in terms of the KLD, the SCLS was worse than the TFLR, 0.132 versus 0.080, respectively, whereas in terms of the JSD, the two models performed equally well, both produced JSD value equal to 0.002.

Table 6: FADN data: estimated regression coefficients of the SCLS and TFLR models.

Estimated coefficients based on the SCLS model										
	Y1	Y2	Y3	Y4	Y5	Y6	Y7	Y8	Y9	Y10
X1	0.9246	0.0000	0.0342	0.0095	0.0000	0.0051	0.0078	0.0120	0.0000	0.0069
X2	0.0000	0.9337	0.0310	0.0000	0.0000	0.0112	0.0163	0.0000	0.0000	0.0078
X3	0.0000	0.0000	1.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000
X4	0.0210	0.0285	0.0063	0.5571	0.0000	0.1653	0.2217	0.0000	0.0000	0.0000
X5	0.0000	0.0000	0.0427	0.0000	0.9229	0.0333	0.0011	0.0000	0.0000	0.0000
X6	0.0000	0.0000	0.0000	0.0000	0.0000	1.0000	0.0000	0.0000	0.0000	0.0000
X7	0.0000	0.0000	0.0649	0.0000	0.0000	0.0000	0.9351	0.0000	0.0000	0.0000
X8	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	1.0000	0.0000	0.0000
X9	0.0232	0.0197	0.0068	0.0025	0.0117	0.0313	0.0115	0.1062	0.7485	0.0385
X10	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	1.0000
Estimated coefficients based on the TFLR model										
	Y1	Y2	Y3	Y4	Y5	Y6	Y7	Y8	Y9	Y10
X1	1.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000
X2	0.0000	0.9995	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0005
X3	0.0000	0.0000	1.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000
X4	0.0000	0.0000	0.0000	0.8653	0.0000	0.0556	0.0791	0.0000	0.0000	0.0000
X5	0.0000	0.0000	0.0000	0.0000	1.0000	0.0000	0.0000	0.0000	0.0000	0.0000
X6	0.0000	0.0000	0.0000	0.0000	0.0000	0.9985	0.0000	0.0000	0.0015	0.0000
X7	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	1.0000	0.0000	0.0000	0.0000
X8	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.9993	0.0007	0.0000
X9	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0392	0.9608	0.0000
X10	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	1.0000

5.3 Education level of father and mother

This data set contains the education level of father and mother in percentages of low (l), medium (m), and high (h) of 31 countries in Europe[14] and the simplicial response is the father's education level, while the simplicial predictor is the mother's educational level. This data set was used by Fiksel et al. (2022) and was chosen on the grounds of illustrating a) the interpretation of the coefficients (see Section 3.2), and b) their 95% confidence regions (see 3.3).

The matrix of the estimated regression coefficients of the SCLS model and of the TFLR model (for comparison purposes) are given below

$$\hat{\mathbf{B}}_{SCLS} = \begin{pmatrix} 0.9014 & 0.0559 & 0.0428 \\ 0.0000 & 0.9409 & 0.0591 \\ 0.0000 & 0.0737 & 0.9263 \end{pmatrix} \text{ and } \hat{\mathbf{B}}_{TFLR} = \begin{pmatrix} 0.9113 & 0.0512 & 0.0375 \\ 0.0000 & 0.9054 & 0.0946 \\ 0.0000 & 0.1415 & 0.8585 \end{pmatrix},$$

where the rows correspond to the low, medium and high educational levels of the mother, whereas the columns indicate the same educational levels for the father. If the percentage of low educated mothers increases (additively) by δ while the percentage of medium educated mothers decreases (additively) by δ, the expected change in the three educational levels of the father is $\delta\,(0.9014 - 0, 0.0559 - 0.9409, 0.0428 - 0.0591) = (0.914\delta, -0.885\delta, -0.0163\delta)$. Figure 3 presents the 95% confidence regions of the three row coefficients. Evidently, there is high uncertainty in the coefficients corresponding to the highly educated women (third

[14]The dataset is available from the *R* package *robCompositions* (Templ et al., 2023).

Table 7: Catalan elections data: yearly average voting proportions by candidate, blank and null votes.

	Voting proportions								
Year	CiU	PSC	PP	IC	ERC	CC	Other	Blank	Null
1980	0.3095	0.1947	0.0157	0.1186	0.1018	0.0000	0.2482	0.0058	0.0058
1984	0.5640	0.2208	0.0745	0.0370	0.0533	0.0000	0.0395	0.0047	0.0062
1988	0.5385	0.2358	0.0503	0.0464	0.0573	0.0000	0.0588	0.0068	0.0060
1992	0.5260	0.2250	0.0540	0.0397	0.1011	0.0000	0.0363	0.0116	0.0064
1995	0.4869	0.2172	0.1022	0.0516	0.1227	0.0000	0.0064	0.0091	0.0038
1999	0.4669	0.3025	0.0734	0.0076	0.1170	0.0000	0.0202	0.0089	0.0035
2003	0.3877	0.2467	0.0875	0.0497	0.2039	0.0000	0.0116	0.0090	0.0039
2006	0.3769	0.2329	0.0792	0.0720	0.1797	0.0125	0.0194	0.0209	0.0065

row coefficients) as depicted by the figure.

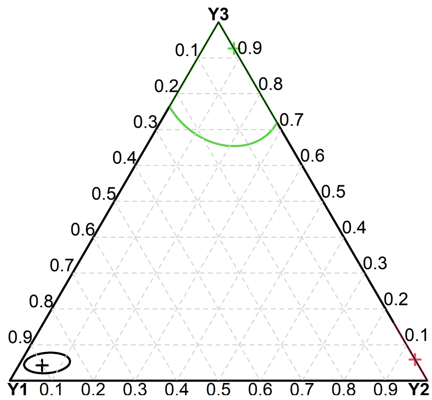


Figure 3: Ecucation level data: confidence regions of the coefficients of the SCLS model: the black refers to the first row coefficients, the red refers to the second row and the green refers to the third row of coefficients.

5.4 Predictive performance of the power transformed SCLS model

The FADN and the education data sets were further used to test the performance of the SCLS model when applied to the power transformed data using Eq. (5) employing the 20 times repeated 10-fold CV protocol. Figure 4 presents the average predictive KLD measures as a function of the α-values. Evidently, this strategy was not proved prosperous for the FADN data, but it was beneficiary in the case of the Education data set.

6 Conclusions

The paper proposed the constrained least squares approach to estimate the parameters of a linear transformation-free regression model for compositional data with compositional predictors. This regression model was first proposed by Fiksel et al. (2022) who, in contrast to this approach, estimated its parameters using the KLD from the observed to the fitted compositional data. This paper showed that extensions to multiple simplicial predictors can straightforwardly be adopted by either approach

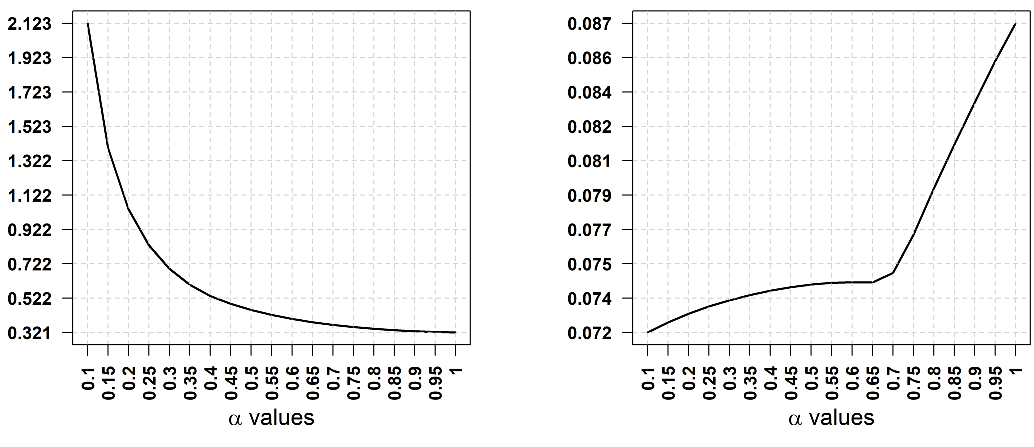
Table 8: Catalan elections data: estimated regression coefficients of the SCLS and TFLR models.

Estimated coefficients based on the SCLS model

	CiU	PSC	PP	IC	ERC	CC	Other	Blank	Null
CiU_{t-1}	0.8326	0.0392	0.0000	0.0000	0.0632	0.0000	0.0390	0.0165	0.0094
PSC_{t-1}	0.0000	0.7031	0.0910	0.1502	0.0481	0.0000	0.0077	0.0000	0.0000
PP_{t-1}	0.0000	0.4776	0.5000	0.0000	0.0224	0.0000	0.0000	0.0000	0.0000
IC_{t-1}	0.1725	0.5814	0.0207	0.0738	0.0000	0.0000	0.1516	0.0000	0.0000
ERC_{t-1}	0.1854	0.0000	0.0679	0.0000	0.7468	0.0000	0.0000	0.0000	0.0000
CC_{t-1}	0.1111	0.1111	0.1111	0.1111	0.1111	0.1111	0.1111	0.1111	0.1111
Other_{t-1}	0.9222	0.0000	0.0778	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000
Blank_{t-1}	0.0000	0.0000	1.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000
Null_{t-1}	1.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000

Estimated coefficients based on the TFLR model

	CiU	PSC	PP	IC	ERC	CC	Other	Blank	Null
CiU_{t-1}	0.8605	0.0357	0.0000	0.0000	0.0548	0.0000	0.0334	0.0101	0.0055
PSC_{t-1}	0.0000	0.7452	0.0711	0.1368	0.0380	0.0000	0.0000	0.0088	0.0000
PP_{t-1}	0.0000	0.4217	0.5342	0.0000	0.0272	0.0000	0.0000	0.0168	0.0000
IC_{t-1}	0.1908	0.5020	0.0000	0.1158	0.0000	0.0000	0.1902	0.0000	0.0012
ERC_{t-1}	0.1499	0.0000	0.0636	0.0000	0.7863	0.0000	0.0000	0.0002	0.0000
CC_{t-1}	0.1111	0.1111	0.1111	0.1111	0.1111	0.1111	0.1111	0.1111	0.1111
Other_{t-1}	0.8211	0.0000	0.1422	0.0000	0.0000	0.0000	0.0277	0.0000	0.0089
Blank_{t-1}	0.0000	0.0000	1.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000
Null_{t-1}	0.1904	0.0000	0.0000	0.2025	0.0000	0.0000	0.2198	0.0803	0.3071



(a) FADN data

(b) Education data

Figure 4: Predictive performance measured by KLD for the α-SCLS model as a function of the α-values.

and further, the categorical predictors case was examined via small simulation studies and the evidence was against the use of this model. The extensive simulation studies comparing both approaches provided evidence that both models perform nearly equally well and there is no clear winner. Finally, the inclusion of a power parameter in the simplicial response provided evidence of increased predictive performance, at the cost of parameter interpretability. Thus, both approaches share similar properties and can be seen as two sides of the same coin. An advantage however of the SCLS over its competitor, the TFLR, is the first's high computational efficiency.

Closing this paper we would like to pose some possible research directions and suggestions. a) Exploitation of the relationship between compositional and directional data. By taking the square root both variables transform into directional data for which spherical regression models exist and in order to ensure that the fitted values will lie within the simplex the folded Kent model (Scealy and Welsh, 2014) may be employed. b) Change of the loss function to the L_1 norm, i.e. constrained minimization of the sum of the absolute errors. c) Investigation of the ensemble learning in the α-SCLS model. Instead of selecting one value of α combine the fitted values of many models produced by different values of α. d) The possibility to add non-linear effects in the model, regardless of the loss function used, to increase the flexibility of the model and escape the assumption of linear relationship.

Appendix

A1 Example of the quadratic programming formulation

Table 12 illustrates an example of the sub-matrices $\mathbf{A}_1^\top$, $\mathbf{A}_2^\top$ and $\mathbf{A}_3^\top$ of the matrix $\mathbf{A}^\top$ and the vector $\mathbf{b}_0$. In this case both simplicial response and predictor variables contain 3 components ($D_r = D_p = 3$).

$$
\mathbf{A}^\top = \begin{pmatrix} \mathbf{A}_1^\top \\ \mathbf{A}_2^\top \\ \mathbf{A}_3^\top \end{pmatrix} = \left(\begin{array}{ccccccccc}
1 & 0 & 0 & 1 & 0 & 0 & 1 & 0 & 0 \\
0 & 1 & 0 & 0 & 1 & 0 & 0 & 1 & 0 \\
0 & 0 & 1 & 0 & 0 & 1 & 0 & 0 & 1 \\
\hline
1 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\
0 & 1 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\
0 & 0 & 1 & 0 & 0 & 0 & 0 & 0 & 0 \\
0 & 0 & 0 & 1 & 0 & 0 & 0 & 0 & 0 \\
0 & 0 & 0 & 0 & 1 & 0 & 0 & 0 & 0 \\
0 & 0 & 0 & 0 & 0 & 1 & 0 & 0 & 0 \\
0 & 0 & 0 & 0 & 0 & 0 & 1 & 0 & 0 \\
0 & 0 & 0 & 0 & 0 & 0 & 0 & 1 & 0 \\
0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 1 \\
\hline
-1 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\
0 & -1 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\
0 & 0 & -1 & 0 & 0 & 0 & 0 & 0 & 0 \\
0 & 0 & 0 & -1 & 0 & 0 & 0 & 0 & 0 \\
0 & 0 & 0 & 0 & -1 & 0 & 0 & 0 & 0 \\
0 & 0 & 0 & 0 & 0 & -1 & 0 & 0 & 0 \\
0 & 0 & 0 & 0 & 0 & 0 & -1 & 0 & 0 \\
0 & 0 & 0 & 0 & 0 & 0 & 0 & -1 & 0 \\
0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & -1
\end{array}\right)^\top \quad \text{and} \quad \mathbf{b}_0 = \begin{pmatrix} 1 \\ 1 \\ 1 \\ 0 \\ 0 \\ 0 \\ 0 \\ 0 \\ 0 \\ 0 \\ 0 \\ 0 \\ -1 \\ -1 \\ -1 \\ -1 \\ -1 \\ -1 \\ -1 \\ -1 \\ -1 \end{pmatrix}
$$

A2 Information (entropy) of the coefficients

Figure A1 shows the negated entropy of data points in the 3-part simplex, $\mathbb{S}^2$. The entropy is maximized (or the negated entropy is maximised) as we move towards the barycentre of the triangle, i.e. the center of the simplex.

A3 A computationally efficient implementation of the linear test of independence

Instead of calling the function *ols.compcomp()* from the *R* package *Compositional* multiple times, a function was created minimizing the necessary to perform computations. The vector $\mathbf{b}_0$ is always the same, and so is the matrix $\mathbf{D}$ since even by permuting the rows of the simplicial predictor $\mathbf{X}$ the cross-product $\mathbf{X}^\top\mathbf{X}$ remains the same. The only thing that changes is the $\mathbf{d}$ vector. Secondly, Eq. (11) was computed excluding the trace of the matrix $\mathbf{Y}\mathbf{Y}^\top$ as this is constant.

A4 Details on the data generation for type II error and discrepancy of the estimated coefficients

Random vectors $\mathbf{x}_i$, for $i = 1, \ldots, n$, were generated from $\text{Dir}\,(1, 1, 1)$, then transformed into $\boldsymbol{\mu}_i = \mathbf{x}_i\mathbf{B}$ and finally random vectors $\mathbf{y}_i$ were generated from $\text{Dir}\,(5\mu_1, \ldots, 5\mu_{D_r})$. The following 4 matrices $\mathbf{B}$ were used.

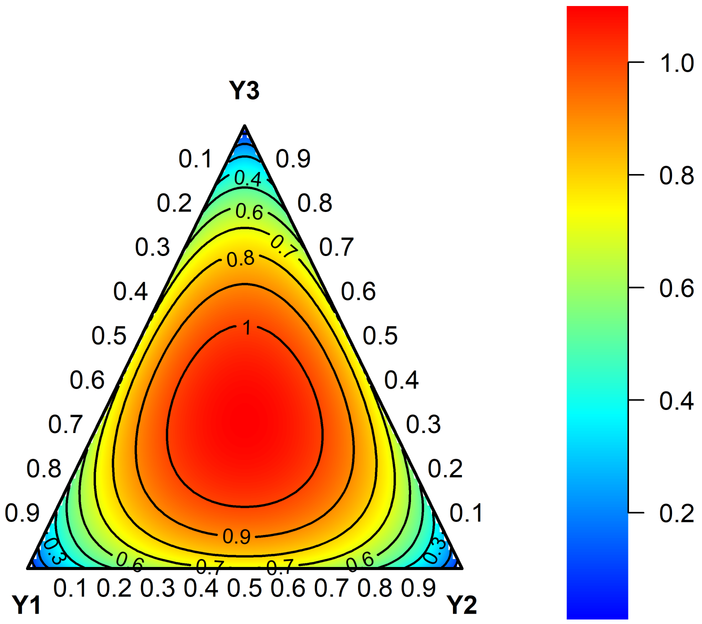


Figure A1: Negated entropy of data points in the simplex.

$$\mathbf{B}_3 = \begin{pmatrix} 0.45 & 0.00 & 0.55 \\ 0.20 & 0.34 & 0.46 \\ 0.76 & 0.01 & 0.23 \end{pmatrix}$$

$$\mathbf{B}_5 = \begin{pmatrix} 0.31 & 0.00 & 0.04 & 0.65 & 0.01 \\ 0.02 & 0.01 & 0.00 & 0.48 & 0.48 \\ 0.28 & 0.02 & 0.64 & 0.06 & 0.00 \end{pmatrix}$$

$$\mathbf{B}_7 = \begin{pmatrix} 0.16 & 0.20 & 0.00 & 0.11 & 0.32 & 0.12 & 0.09 \\ 0.63 & 0.08 & 0.00 & 0.09 & 0.10 & 0.08 & 0.01 \\ 0.10 & 0.24 & 0.20 & 0.12 & 0.03 & 0.01 & 0.30 \end{pmatrix}$$

$$\mathbf{B}_{10} = \begin{pmatrix} 0.25 & 0.00 & 0.01 & 0.09 & 0.01 & 0.00 & 0.24 & 0.14 & 0.00 & 0.26 \\ 0.44 & 0.10 & 0.18 & 0.02 & 0.01 & 0.00 & 0.09 & 0.07 & 0.00 & 0.10 \\ 0.34 & 0.03 & 0.00 & 0.14 & 0.17 & 0.00 & 0.04 & 0.00 & 0.19 & 0.09 \end{pmatrix}$$

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

Alenazi, A. (2019). Regression for compositional data with compositional data as predictor variables with or without zero values. *Journal of Data Science 17*(1), 219–237.

Chen, J., X. Zhang, and S. Li (2017). Multiple linear regression with compositional response and covariates. *Journal of Applied Statistics 44*(12), 2270–2285.

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.

Fiksel, J. and A. Datta (2021). *codalm: Transformation-Free Linear Regression for Compositional Outcomes and Predictors.* R package version 0.1.2.

Fiksel, J., S. Zeger, and A. Datta (2022). A transformation-free linear regression for compositional outcomes and predictors. *Biometrics 78*(3), 974–987.

Goldfarb, D. and A. Idnani (1983). A numerically stable dual method for solving strictly convex quadratic programs. *Mathematical Programming 27*(1), 1–33.

Han, H. and K. Yu (2022). Partial linear regression of compositional data. *Journal of the Korean Statistical Society 51*(4), 1090–1116.

Higham, N. J. (2002). Computing the nearest correlation matrix—a problem from finance. *IMA journal of Numerical Analysis 22*(3), 329–343.

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

James, G. M., C. Paulson, and P. Rusmevichientong (2019). Penalized and constrained optimization: an application to high-dimensional website advertising. *Journal of the American Statistical Association 115*(529), 107–122.

Jolliffe, I. T. (2005). *Principal Component Analysis.* New York: Springer-Verlag.

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.

Liew, C. K. (1976). Inequality constrained least-squares estimation. *Journal of the American Statistical Association 71*(355), 746–751.

Lu, S., W. Wang, and R. Guan (2024). Kent feature embedding for classification of compositional data with zeros. *Statistics and Computing 34*(2), 1–17.

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.

Mattas, K., M. Tsagris, and V. Tzouvelekas (2024). Using synthetic farm data to estimate individual nitrate leaching levels. Technical report, Department of Economics, University of Crete.

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.

Owen, A. B. (2001). *Empirical likelihood*. Chapman and Hall/CRC.

Papadakis, M., M. Tsagris, S. Fafalios, and M. Dimitriadis. (2023). *Rfast2: A Collection of Efficient and Extremely Fast R Functions II*. R package version 0.1.5.1.

Pison, G., A. Struyf, and P. J. Rousseeuw (1999). Displaying a clustering with CLUSPLOT. *Computational Statistics & Data Analysis 30*(4), 381–392.

Scealy, J. and A. H. Welsh (2014). Fitting Kent models to compositional data with small concentration. *Statistics and Computing 24*, 165–179.

Templ, M., K. Hron, P. Filzmoser, K. Facevicova, P. Kynclova, J. Walach, V. Pintar, J. Chen, D. Miksova, B. Meindl, A. Menafoglio, A. Di Blasi, F. Pavone, N. Stefelova, G. Zeni, and R. Wiederkehr (2023). *robCompositions: Compositional Data Analysis*. R package version 2.4.1.

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.

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., A. Alenazi, and C. Stewart (2023). Flexible non-parametric regression models for compositional response data with zeros. *Statistics and Computing 33*(5), 106.

Tsagris, M., G. Athineou, A. Alenazi, and C. Adam (2024). *Compositional: Compositional Data Analysis*. R package version 6.7.

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., S. Preston, and A. T. A. Wood (2017). Nonparametric hypothesis testing for equality of means on the simplex. *Journal of Statistical Computation and Simulation 87*(2), 406–422.

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

Tsagris, M. and C. Stewart (2022). *A Review of Flexible Transformations for Modeling Compositional Data*, pp. 225–234. Cham: Springer International Publishing.

Turlach, B. A., A. Weingessel, and C. Moler (2019). *quadprog: Functions to Solve Quadratic Programming Problems*. R package version 1.5-8.

Van Aelst, S. and P. Rousseeuw (2009). Minimum volume ellipsoid. *Wiley Interdisciplinary Reviews: Computational Statistics 1*(1), 71–82.

Wang, H., J. Meng, and M. Tenenhaus (2009). Regression modelling analysis on compositional data. In *Handbook of Partial Least Squares: Concepts, Methods and Applications*, pp. 381–406. Springer.

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

Wets, R. J. B. (1991). Constrained estimation: Consistency and asymptotics. *Applied Stochastic Models and Data Analysis 7*(1), 17–32.