A Regression Model for Compositional Data Based on the Shifted-Dirichlet Distribution

G.S. Monti, G. Mateu-Figueras, V. Pawlowsky-Glahn and J.J. Egozcue

Abstract Using an approach based on the Aitchison geometry of the simplex, a Shifted-Dirichlet covariate model is obtained. Allowing the parameters to change linearly with a set of covariates, their effects on the relative contributions of different components in a composition are assessed. An application of this model to sedimentary petrography is given.

Keywords Dirichlet regression · Simplicial regression · Model selection

1 Introduction

Compositional data are vectors of parts of some whole which carry relative information. They are frequently represented as proportions or percentages, which are subject to a constant sum, κ, i.e., $\kappa = 1$ or $\kappa = 100$. Their sample space is then represented by the simplex, denoted by

G.S. Monti (✉)
Department of Economics, Management and Statistics, University of Milano-Bicocca, Milano, Italy
e-mail: gianna.monti@unimib.it

G. Mateu-Figueras · V. Pawlowsky-Glahn
Department of Computer Science, Applied Mathematics and Statistics, University of Girona, Girona, Spain
e-mail: gloria.mateu@udg.edu

V. Pawlowsky-Glahn
e-mail: vera.pawlowsky@udg.edu

J.J. Egozcue
Department of Civil and Environmental Engineering, Technical University of Catalonia, Barcelona, Spain
e-mail: juan.jose.egozcue@upc.edu

J.A. Martín-Fernández and S. Thió-Henestrosa (eds.), *Compositional Data Analysis*, Springer Proceedings in Mathematics & Statistics 187,
DOI 10.1007/978-3-319-44811-4_9

$$\mathscr{S}^D = \{\mathbf{x} = (x_1, \ldots, x_D),\ x_i > 0,\ \sum_{i=1}^{D} x_i = \kappa\}.$$

Compositional data occur in many applied fields: from geology and biology to election forecast, from medicine and psychology to economic studies.

We recall briefly the essential elements of simplicial algebra, as it will be used later. For any vector of D strictly positive real components,

$$\mathbf{z} = (z_1, \ldots, z_D) \in \mathbb{R}_+^D \quad z_i > 0, \text{ for all } i = 1, \ldots, D,$$

the *closure* operation of $\mathbf{z}$ is defined as

$$\mathscr{C}(\mathbf{z}) = \left(\frac{\kappa\ z_1}{\sum_{i=1}^{D} z_i}, \ldots, \frac{\kappa\ z_D}{\sum_{i=1}^{D} z_i} \right) \in \mathscr{S}^D . \tag{1}$$

where κ is the sum of the components, i.e., the constraint.

The two basic operations required for a vector space structure of the simplex are *perturbation*: given two compositions $\mathbf{x}$ and $\mathbf{y} \in \mathscr{S}^D$,

$$\mathbf{x} \oplus \mathbf{y} = \mathscr{C}\left(x_1 y_1, \ldots, x_D y_D\right), \tag{2}$$

and *powering*: given a composition $\mathbf{x} \in \mathscr{S}^D$ and a scalar $\alpha \in \mathbb{R}$,

$$\alpha \odot \mathbf{x} = \mathscr{C}\left(x_1^\alpha, \ldots, x_D^\alpha\right) . \tag{3}$$

Furthermore, an *inner product* $\langle \cdot, \cdot \rangle_a$ is defined as

$$\langle \mathbf{x}, \mathbf{y} \rangle_a = \sum_{i=1}^{D} \ln \frac{x_i}{\mathrm{g_m}(\mathbf{x})} \ln \frac{y_i}{\mathrm{g_m}(\mathbf{y})}, \tag{4}$$

where $\mathrm{g_m}(\mathbf{x})$ denotes the geometric mean of the components of $\mathbf{x}$ [4, 22]. As shown in Pawlowsky-Glahn and Egozcue [22] the simplex $(\mathscr{S}^D, \oplus, \odot, \langle \cdot, \cdot \rangle_a)$ has a $(D-1)$-dimensional real Euclidean vector space structure called *simplicial* or *Aitchison geometry*.

Let $(\mathbf{e}_1, \mathbf{e}_2, \ldots, \mathbf{e}_{D-1})$ be an orthonormal basis of the simplex and consider the $(D-1) \times D$ matrix $\boldsymbol{\Psi}$ which rows are $\Psi_i = \mathrm{clr}(\mathbf{e}_i)$, $(i = 1, \ldots, D-1)$. Note that clr is the centered log-ratio transformation, a function from $\mathscr{S}^D$ to $\mathbb{R}^D$ defined as

$$\mathrm{clr}(\mathbf{x}) = \left(\log \frac{x_1}{\mathrm{g_m}(\mathbf{x})}, \ldots, \log \frac{x_D}{\mathrm{g_m}(\mathbf{x})} \right),$$

where $g_m(\mathbf{x})$ is the geometric mean of the D components of $\mathbf{x}$. The $\boldsymbol{\Psi}$ matrix is called *contrast matrix* associated with the orthonormal basis $(\mathbf{e}_1, \ldots, \mathbf{e}_{D-1})$. Each row is called a (log)contrast.

The *isometric log-ratio transformation*, ilr for short, of $\mathbf{x}$ is the function $\text{ilr} : \mathscr{S}^D \to \mathbb{R}^{D-1}$, which assigns coordinates $\mathbf{x}^\star$, with respect to the given basis, to the composition $\mathbf{x}$. The vector $\mathbf{x}^\star$ contains the $D-1$ ilr-coordinates of $\mathbf{x}$ in a Cartesian coordinate system. The inverse of the ilr-transformation is denoted by ilr^{-1}. The function ilr is an isometry of vector spaces. The ilr transformation is computed as a simple matrix product:

$$\mathbf{x}^\star = \text{ilr}(\mathbf{x}) = \ln(\mathbf{x})\boldsymbol{\Psi}' .$$

Inversion of ilr, i.e., recovering the composition from its coordinates, is given by

$$\mathbf{x} = \text{ilr}^{-1}(\mathbf{x}^\star) = \mathscr{C}\left(\exp(\boldsymbol{\Psi}\ \mathbf{x}^\star)\right) .$$

Given an orthonormal basis of the simplex, any composition $\mathbf{x} \in \mathscr{S}^D$ can be expressed as a linear combination,

$$\begin{aligned}\mathbf{x} &= (x_1^\star \odot \mathbf{e}_1) \oplus (x_2^\star \odot \mathbf{e}_2) \oplus \cdots \oplus (x_{D-1}^\star \odot \mathbf{e}_{D-1}) \\ &= \bigoplus_{i=1}^{D-1} (x_i^\star \odot \mathbf{e}_i),\end{aligned}$$

where the symbol $\bigoplus$ represents repeated perturbation. The coefficients of the linear combination, for a fixed basis, are uniquely determined, given that in a Euclidean space any point can always be represented in a unique way by its coordinates with respect to an orthonormal basis. Once an orthonormal basis has been chosen, all standard statistical methods can be applied to coordinates and transferred to the simplex preserving their properties [15].

A natural measure on $\mathscr{S}^D$, called *Aitchison measure*, can be defined using orthonormal coordinates [21, 23], that is, the Aitchison measure of a subset on the simplex is the Lebesgue measure of the subset in the space of orthonormal coordinates. This measure is compatible with the *Aitchison geometry* and is absolutely continuous with respect to the Lebesgue measure on the D-dimensional real space. The relationship between them is $\sqrt{D}x_1x_2\ldots x_D$. The change of the reference measure has some important implications, for example to compute the expected value (see [16] for an in-depth discussion).

Historically, there are essentially two different approaches to regression models which relate a compositional response variable with a system of covariates: Simplicial regression and Dirichlet regression. The former is based on the Aitchison's theoretical result that if a compositional vector follows an additive logistic normal distribution, the log-ratio transformed vector will follow a normal distribution [2, 3, 8]; the latter follows the *stay-in-the-simplex* approach. It assumes that the response variable follows a Dirichlet distribution whose parameters are a log-linear function

of a set of covariates [6, 12, 14, 17]. Other solutions, present in the literature but less used, involve models based on the generalized Liouville distribution [25].

The Scaled-Dirichlet distribution is an extension of the Dirichlet one. Given that we work here with the Aitchison geometry of the simplex, and that within this framework it is a perturbation of a standard Dirichlet [18], we will refer hereafter to it as the Shifted-Dirichlet distribution. The reason to change this terminology is twofold. On the one hand, working within the Aitchison geometry implies a change of the reference measure; on the other hand, scaling in this geometry is achieved using a power transformation, which allows another extension already studied in Monti et al. [19]. In summary, the name of the distribution indicates the sample space of the corresponding random vector and its structure. For the Scaled-Dirichlet distribution this is the simplex as a subset of real space with the induced Euclidean geometry, while for the Shifted-Dirichlet distribution it is the simplex as a Euclidean space endowed with the Aitchison geometry. Although in the first case the Lebesgue reference measure is used, and in the second the Aitchison measure, the probability assigned to any measurable subset of the simplex is the same.

The Shifted-Dirichlet covariate model is an extension of the Dirichlet one, based on the algebraic geometric structure of the simplex. The assumption is that $\mathbf{x} = (x_1, \ldots, x_D)$ is a compositional response vector, with D components having a Shifted-Dirichlet distribution, in which the parameters $\boldsymbol{\alpha}$ are allowed to change with a set of covariates.

The paper is structured as follows. Section 2 defines the two existing approaches: Simplicial regression and Dirichlet regression. Section 3 gives a brief overview of the Shifted-Dirichlet distribution and describes the Shifted-Dirichlet covariate model, dealing with the issue of parameter estimation. Section 4 presents an example of application of the proposed regression model to sedimentary petrography, in particular bulk petrography and heavy-mineral data of Pleistocene sands (Regione Lombardia cores; Po Plain); this dataset is described in Garzanti et al. [11].

2 Regression Models for Compositional Response Variable

2.1 Simplicial Regression

Linear regression with compositional response variable can be stated as follows. A compositional sample of n independent observations, denoted by $\mathbf{x}_1, \ldots, \mathbf{x}_n$, is available. Each data point, $\mathbf{x}_j$, $(j = 1, \ldots, n)$ is associated with one or more external variables or covariates grouped in the vector $\mathbf{s}_j = (s_{j0}, s_{j1}, \ldots, s_{jm}, \ldots, s_{jp})$, where $s_{j0} = 1$ by convention.

The basic idea of Simplicial regression [8] relies on the principle of working on coordinates: once a basis is chosen, the associated ilr coordinates are computed and the classical regression of the ilr coordinates on the covariates is performed. Through

the inverse ilr-transformation, the back-transformed coefficient vectors, as well as predictions and confidence intervals are obtained.

The general model can be expressed as

$$\hat{\mathbf{x}}(\mathbf{s}) = (s_0 \odot \boldsymbol{\delta}_0) \oplus (s_1 \odot \boldsymbol{\delta}_1) \oplus \cdots \oplus \left(s_p \odot \boldsymbol{\delta}_p\right) = \bigoplus_{m=0}^{p} s_m \odot \boldsymbol{\delta}_m \,. \tag{5}$$

Note that there are $p + 1$ coefficient vectors δ_m, as many as covariates, and that they are vectors with $(D - 1)$ components, as many as coordinates. The goal of estimating the coefficients $\boldsymbol{\delta}$ of a curve or surface in $\mathscr{S}^D$ is solved by translating it into a $(D - 1)$ least square problem, i.e., for each coordinate

$$\hat{\mathbf{x}}_i^{\star}(\mathbf{s}) = \delta_{0i}^{\star} s_0 + \delta_{1i}^{\star} s_1 + \cdots + \delta_{pi}^{\star} s_p, \quad i = 1, \ldots, D - 1, \tag{6}$$

where $\boldsymbol{\delta}_m^{\star} = \left(\delta_{m1}^{\star}, \ldots, \delta_{m,D-1}^{\star}\right)$ is the coordinate vector associated with $\boldsymbol{\delta}_m$. In the case of simple regression $m = 1$ and $\mathbf{s} = s$, which is a straight-line in the simplex.

2.2 Dirichlet Regression

The Dirichlet distribution is one of the well known probability models suitable for random compositions. A random vector $\mathbf{X} = (X_1, \ldots, X_D) \in \mathscr{S}^D$ has a Dirichlet distribution, indicated by $\mathbf{X} \sim \mathscr{D}^D(\boldsymbol{\alpha})$, with $\boldsymbol{\alpha} = (\alpha_1, \ldots, \alpha_D) \in \mathbb{R}_+^D$, when its density function (with respect to the Aitchison measure) is

$$f(\mathbf{x}; \boldsymbol{\alpha}) = \frac{\sqrt{D}\,\Gamma(\alpha_+)}{\prod_{i=1}^{D} \Gamma(\alpha_i)} \prod_{i=1}^{D} x_i^{\alpha_i}, \tag{7}$$

where $\alpha_+ = \sum_{i=1}^{D} \alpha_i$, and Γ denotes the gamma function [18]. The Dirichlet distribution has D parameters α_i, which are assumed to be positive. Note that the density (7) is obtained by changing the measure to a Dirichlet density with respect to the Lebesgue measure.

In the Dirichlet regression model the α_i parameters are reparameterized in terms of explanatory variables and coefficients through an exponential function as described in Eq. (12). The log-likelihood of the reparameterized Dirichlet distribution can be optimized via an iterative method such as the Newton–Raphson algorithm.

When the variable of interest is continuous and restricted to the unit interval (0, 1), i.e., when $D = 2$, the Dirichlet regression is called Beta regression [10].

3 Shifted-Dirichlet Covariate Model

3.1 Shifted-Dirichlet Distribution

One of the generalizations of the Dirichlet distribution is the Shifted-Dirichlet distribution. A random vector $\mathbf{X} \in \mathscr{S}^D$ has a Shifted-Dirichlet distribution with parameters $\boldsymbol{\alpha}$ and $\boldsymbol{\beta} = (\beta_1, \ldots, \beta_D) \in \mathscr{S}^D$ if its density function is

$$f(\mathbf{x}, ; \boldsymbol{\alpha}, \boldsymbol{\beta}) = \frac{\Gamma(\alpha_+)\sqrt{D}}{\prod_{i=1}^D \Gamma(\alpha_i)} \frac{\prod_{i=1}^D \left(\frac{x_i}{\beta_i}\right)^{\alpha_i}}{\left(\sum_{i=1}^D \frac{x_i}{\beta_i}\right)^{\alpha_+}}, \tag{8}$$

The density (8) is expressed with respect to the Aitchison probability measure [21]. See Monti et al. [18] for a detailed discussion about the reasons and implications to use the Aitchison measure. This distribution will be denoted by $\mathbf{X} \sim \mathscr{SD}^D(\boldsymbol{\alpha}, \boldsymbol{\beta})$.

The number of parameters of this model is $2D - 1$, since $\boldsymbol{\beta} \in \mathscr{S}^D$. The Shifted-Dirichlet distribution can be obtained by normalizing a vector of D independent, scaled (in the Euclidean geometry of real space), gamma r.v.s $W_i \sim Ga(\alpha_i, \beta_i)$, $i = 1, 2, \ldots, D$; i.e., if $\mathbf{X} = \mathscr{C}(\mathbf{W})$, with $\mathbf{W} = (W_1, \ldots, W_D) \in \mathbb{R}_+^D$, then $\mathbf{X} \sim \mathscr{SD}^D(\boldsymbol{\alpha}, \boldsymbol{\beta})$ [18]. For this reason, in the literature, when working with the Lebesgue reference measure, the distribution is called Scaled-Dirichlet. This distribution can also be obtained as a perturbed random composition with a Dirichlet density. Recall that perturbation is, in the Aitchison geometry of the simplex, a shift. Therefore, here it is called Shifted-Dirichlet distribution, understanding that the density is expressed with respect to the Aitchison measure.

Indeed, let $\widetilde{\mathbf{X}} \sim \mathscr{D}^D(\boldsymbol{\alpha})$ be a random composition defined in $\mathscr{S}^D$, and let $\boldsymbol{\beta} \in \mathscr{S}^D$ be a composition. The random composition $\mathbf{X} = \ominus\boldsymbol{\beta} \oplus \widetilde{\mathbf{X}}$ has a $\mathscr{SD}^D(\boldsymbol{\alpha}, \boldsymbol{\beta})$ distribution (note that $\ominus$ is the inverse operation of $\oplus$). Observe that using the Aitchison measure and geometry, $\boldsymbol{\beta}$ can be interpreted as a parameter of location, instead of as a measure of scale. The expected value of $\mathbf{X} \sim \mathscr{SD}^D(\boldsymbol{\alpha}, \boldsymbol{\beta})$ with respect to the Aitchison measure is

$$\mathrm{E}_a(\mathbf{X}) = \ominus\boldsymbol{\beta} \oplus \mathrm{E}_a(\widetilde{\mathbf{X}}), \tag{9}$$

where $\mathrm{E}_a(\widetilde{\mathbf{X}})$ is the expected value of a Dirichlet composition with respect to the Aitchison measure

$$\mathrm{E}_a(\widetilde{\mathbf{X}}) = \mathscr{C}(e^{\psi(\alpha_1)}, \ldots, e^{\psi(\alpha_D)}), \tag{10}$$

with ψ the digamma function. The metric variance of $\mathbf{X}$ coincides with the metric variance of a Dirichlet composition, because this measure of dispersion is invariant under perturbation

$$\mathrm{Mvar}(\mathbf{X}) = \frac{D-1}{D}(\psi'(\alpha_1), \ldots, \psi'(\alpha_D)), \tag{11}$$

with ψ' the trigamma function [1].

3.2 Shifted-Dirichlet Regression

Given a sample of n independent compositional observations $(\mathbf{x}_1, \ldots, \mathbf{x}_n)$, we hypothesize that each observation $\mathbf{x}_j$ follows a conditional Shifted-Dirichlet distribution, given a set of covariates. Polynomial regression on a covariate s is included as a particular case taking $s_{jm} = s_j^m$.

In order to incorporate the covariate effects into the model [6, 14], we reparameterize each parameter α_i of the density written in Eq. (8) in terms of covariates and regression coefficients via the following log-linear model

$$\alpha_{ij} = \alpha_{ij}(\mathbf{s}_j) = \exp\left\{\sum_{m=0}^{p} \delta_{im} s_{jm}\right\}, \tag{12}$$

where $\mathbf{s}_j$ is the covariate vector recorded on the j-th observed composition ($j = 1, \ldots, n$), and δ_{im} are the coefficients for the m-th covariate. The parameter δ_{im} theoretically can vary by component, and the covariates may or may not be the same set of explanatory variables for each α_{ij}. We augment each vector $\mathbf{s}_j$ with 1 as first position for notation simplicity. Thus, given a sample of independent compositional observations of size n, $\mathbf{x}_1, \ldots, \mathbf{x}_j, \ldots, \mathbf{x}_n$ the log-likelihood function for the reparameterized Shifted-Dirichlet, given the covariates $\mathbf{s}$ and ignoring the constant part that does not involve the parameters, is equal to

$$\begin{aligned} l(\boldsymbol{\beta}, \boldsymbol{\delta}|\mathbf{x}, \mathbf{s}) = \sum_{j=1}^{n} \Bigg\{ & \log \Gamma\left(\sum_{i=1}^{D} \exp\left\{\sum_{m=0}^{p} \delta_{im} s_{jm}\right\}\right) - \sum_{i=1}^{D} \log \Gamma\left(\exp\left\{\sum_{m=0}^{p} \delta_{im} s_{jm}\right\}\right) \\ & - \sum_{i=1}^{D} \exp\left\{\sum_{m=0}^{p} \delta_{im} s_{jm}\right\} \log \beta_i + \sum_{i=1}^{D} \exp\left\{\sum_{m=0}^{p} \delta_{im} s_{jm}\right\} \log x_{ij} \\ & - \left(\sum_{i=1}^{D} \exp\left\{\sum_{m=0}^{p} \delta_{im} s_{jm}\right\}\right) \log\left(\sum_{i=1}^{D} \frac{x_i}{\beta_i}\right) \Bigg\}. \end{aligned} \tag{13}$$

Equation (13) can be estimated using the maximum likelihood method via some optimization algorithm, e.g., the Newton–Raphson algorithm. The choice of the starting values for the algorithm is of fundamental importance to get fast convergence.

For the Dirichlet regression in Hijazi and Jernigan [14] a method based on resampling from the original data is proposed; for each resample a Dirichlet model with constant parameters is fitted and the mean of the corresponding covariates is computed. After that, D models of the form $\sum_{m=0}^{p} \delta_{im} s_{jm}$ are fitted by least squares. The fitted coefficients $\hat{\delta}_{im}$ are used as starting values. For the Shifted-Dirichlet covariate model we have followed the same principle; as starting point for the vector $\boldsymbol{\beta}$ we have chosen the closed geometric mean of the components of $\mathbf{x}$ given by

$$\mathbf{g}(\mathbf{x}) = \mathscr{C}\left(\left(\prod_{j=1}^{n} x_{1j}\right)^{1/n}, \ldots, \left(\prod_{j=1}^{n} x_{Dj}\right)^{1/n}\right), \tag{14}$$

Model selection is performed by testing

$$H_0 : \delta_{im} = 0 , \tag{15}$$

for some pair $(i, m), i = 1, \ldots, D$ and $m = 1, \ldots, p$. For it, the traditional likelihood ratio test is implemented.

4 Example from Sedimentology

4.1 Data Description

In Garzanti et al. [11] the authors studied the paleogeographic and paleodrainage changes during Pleistocene glaciations of Po Plain by compositional signatures of Pleistocene sands. In particular we consider here Cilavegna and Ghedi cores of Regione Lombardia, with 18 and 19 compositional observations, respectively. In this section we compare the above-mentioned approaches to regression with a composition as dependent variable. The goal is to model the effect of the depth covariate on compositional signatures of Pleistocene sands taking the fact into account that the cores may have separate effects on the response. The three compositional parts are: Q (= quartz), F (= feldspar), and L (= lithic grains) represented in the usual ternary diagram in Fig. 1.

4.2 Estimated Models Comparison

For each of the three components we have fitted a regression model considering the depth as covariate, including in the model one dummy variable representing the core provenance (0 for Cilavegna and 1 for Ghedi), as well as the interaction term. The categorical covariate core has been included to account for variation in proportions that is a function of a group-specific factor. The Dirichlet and Shifted-Dirichlet covariate models are expressed with respect to the Aitchison measure.

Tables 1 and 2 summarize the estimated regression coefficients, together with results from the inference for the Dirichlet and Shifted-Dirichlet covariate model, respectively.

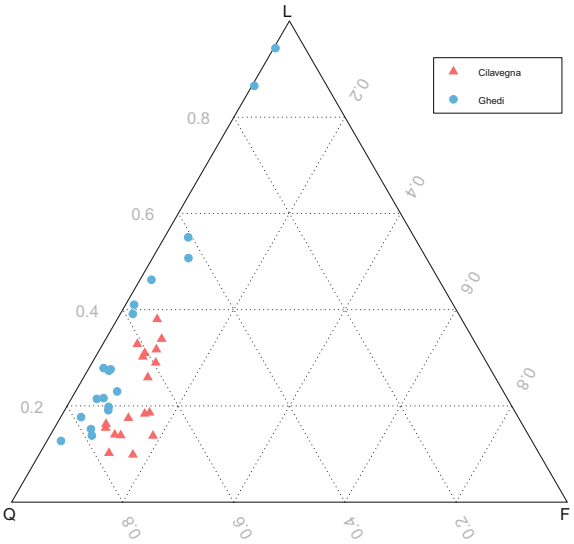

Fig. 1 Quaternary Po Plain sediments: ternary plot. Points are distinguished by core

Table 1 Regression output of the Dirichlet covariate model for Quaternary Po Plain sediments

Regressors	Coefficient	S.E.	z value	Pr(> |z|)
	δ-coefficients for variable Q (= quartz)			
(Intercept)	3.8695	0.8536	4.5329	0.0000
Depth	0.0027	0.0063	0.4304	0.6669
Core	−4.6905	0.5509	−8.5137	0.000
Depth ∗ core	0.0243	0.0038	6.4577	0.0000
	δ-coefficients for variable F (= feldspar)			
(Intercept)	2.2494	0.7975	2.8205	0.0048
Depth	0.0015	0.0058	0.2495	0.8030
Core	−4.3685	0.6012	−7.2662	0.0000
Depth ∗ core	0.0190	0.0039	4.9164	0.0000
	δ-coefficients for variable L (= lithic grains)			
(Intercept)	2.0708	0.7773	2.6641	0.0077
Depth	0.0091	0.0058	1.5794	0.1143
Core	−2.3527	0.3583	−6.5655	0.0000
Depth ∗ core	0.0076	0.0021	3.5803	0.0003

Table 2 Regression output of the Shifted-Dirichlet covariate model for Quaternary Po Plain sediments

	δ-coefficients for variable Q (= quartz)			
Regressors	Coefficient	S.E.	z value	Pr(> |z|)
(Intercept)	3.1275	0.7534	4.1510	0.0000
Depth	0.0035	0.0053	0.6649	0.5061
Core	−3.5720	0.8824	−4.0481	0.0001
Depth ∗ core	0.0171	0.0069	2.4658	0.0137
	δ-coefficients for variable F (= feldspar)			
Regressors	Coefficient	S.E.	z value	Pr(> |z|)
(Intercept)	3.2542	0.7161	4.5442	0.0000
Depth	0.0023	0.0050	0.4549	0.6492
Core	−4.5289	0.7993	−5.6664	0.0000
Depth ∗ core	0.0187	0.0064	2.9332	0.0034
	δ-coefficients for variable L (= lithic grains)			
Regressors	Coefficient	S.E.	z value	Pr(> |z|)
(Intercept)	1.4002	0.6546	2.1390	0.0324
Depth	0.0098	0.0049	2.0100	0.0444
Core	−1.2044	0.8589	−1.4022	0.1608
Depth ∗ core	0.0004	0.0068	0.0516	0.9589
	β-coefficients			
Q	0.4693	0.1362	3.4466	0.0006
F	0.0789	0.0397	1.9895	0.0467

Table 3 Model fit statistics for the nested Shifted-Dirichlet covariate models, where $\Delta G^2 = -2 \log L$(reduced model − current model)

Criterion	Intercept only	Depth covariate	Full model
AIC	210.1447	187.4468	122.5
BIC	218.1993	200.3341	145.053
logL	−100.0724	−85.7233	−47.25
df	5	8	14
ΔG^2		28.697	76.947
Pr > ChiSq		<0.0001	<0.0001

Table 4 Regression output for the first and second coordinate of Simplicial regression for the Quaternary Po Plain sediments

Regressors	Coefficient	S.E.	z value	Pr(> |z|)
	$x_1^\star$ *coordinate*			
(Intercept)	−1.1867	0.1264	−9.3862	0.0000
Depth	−0.0008	0.0011	−0.7209	0.4760
Core	−1.0654	0.1886	−5.6487	0.0000
Depth ∗ core	0.0032	0.0015	2.1193	0.0417
	$x_2^\star$ *coordinate*			
(Intercept)	−0.8479	0.2903	−2.9212	0.0062
Depth	0.0061	0.0024	2.5012	0.0175
Core	2.7908	0.4330	6.4454	0.0000
Depth ∗ core	−0.0174	0.0035	−4.9827	0.0000

The p-value of the likelihood ratio tests to compare the intercept-only model (e.g., no predictors) with the fitted Shifted Dirichlet covariate model is essentially zero (<0.0001), which provides evidence against the reduced model in favor of the current model, as well as the model with only depth as covariate (see Table 3).

In Table 2 it can be seen that the z-values for the first two components are highly significant, implying that the use of the dummy variable is important; the models appear to have a definite nonzero slope. This consideration is confirmed by the fitted regression lines reported in Fig. 2.

Akaike information criterion (AIC) and Bayesian information criterion (BIC) are usually used to compare adequacy of models of the same family, the preferred model is the one with the minimum AIC value or BIC value. For the Dirichlet regression we obtained that $AIC = 131.2837$ and $BIC = 150.6148$, while, for the Shifted-Dirichlet regression the two measures are $AIC = 122.45$ and $BIC = 145.053$ (see Full model in Table 3). Therefore we can conclude that the improvement of the model compensates the additional parameters in the Shifted-Dirichlet model.

In order to apply the Simplicial regression, ilr coordinates of the Quaternary Po Plain sediment dataset are computed (Table 4). The canonical basis in the clr plane was used as ilr transform, so that, the two coordinates or balances are expressed as:

$$x_1^\star = \frac{1}{\sqrt{2}} \log \frac{x_2}{x_1}\,, \quad x_2^\star = \frac{1}{\sqrt{6}} \log \frac{x_3^2}{x_1\, x_2}\,. \tag{16}$$

Predictions of the three coordinates can be back-transformed with the inverse ilr, to obtain a prediction of the proportions themselves.

In order to assess the adequacy of the different regression approaches, we examine some goodness of fit measures. One suitable determination coefficient for the regression model to evaluate the proportion of explained variation in the compositions by the covariate is connected with the total variability [2, 13], based on the variation

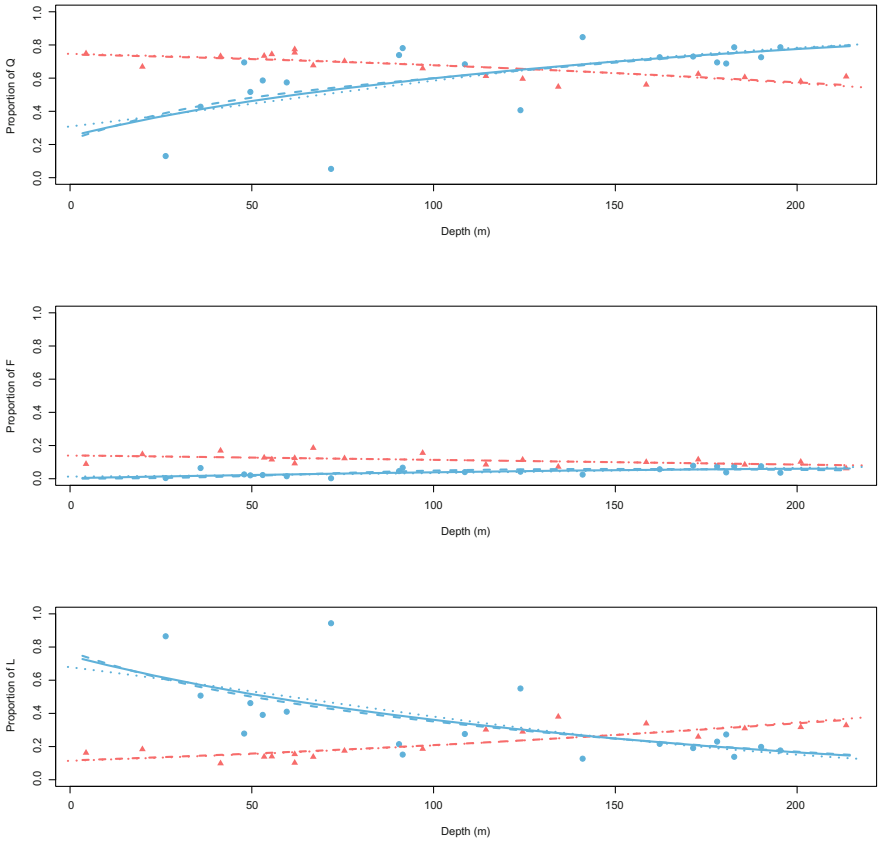

Fig. 2 Observed and fitted compositions for Quaternary Po Plain sediments using the three models for each level of the core variable (Shifted-Dirichlet covariate model: *solid line*, Dirichlet covariate model: *dashed line*; Simplicial regression: *dotted line*). *Red colors* refers to Cilavegna core data and *blue color* refers to Ghedi core data

matrix of the transformed log-ratio data,

$$\mathbf{T}(\mathbf{x}) = [t_{ir}] = \left[\text{var}\left(\ln \frac{x_i}{x_r} \right) \right] \quad i, r = 1, \ldots, D. \tag{17}$$

Each element t_{ir} is the usual variance of the log ratio of parts i and r. Aitchison's total variability measure totvar($\mathbf{x}$), a measure of global dispersion of a compositional sample, is defined as

$$\text{totvar}(\mathbf{x}) = \frac{1}{2D} \sum_{i,r} \text{var}\left(\ln \frac{x_i}{x_r} \right) = \frac{1}{2D} \sum_{i,r} t_{ir}\,, \tag{18}$$

The determination coefficient R_T^2 is defined as

$$R_T^2 = \frac{\text{totvar}(\hat{\mathbf{x}})}{\text{totvar}(\mathbf{x})}; \tag{19}$$

it compares the total variability of the observed with the fitted data.

Table 5 Goodness of fit measures for the three different regression models

	R_T^2	R_A^2	KL-div
Dirichlet regression	0.6914	0.5624	1.6472
Shifted-Dirichlet regression	0.5733	0.5965	1.6209
Simplicial regression	0.5907	0.5907	1.657

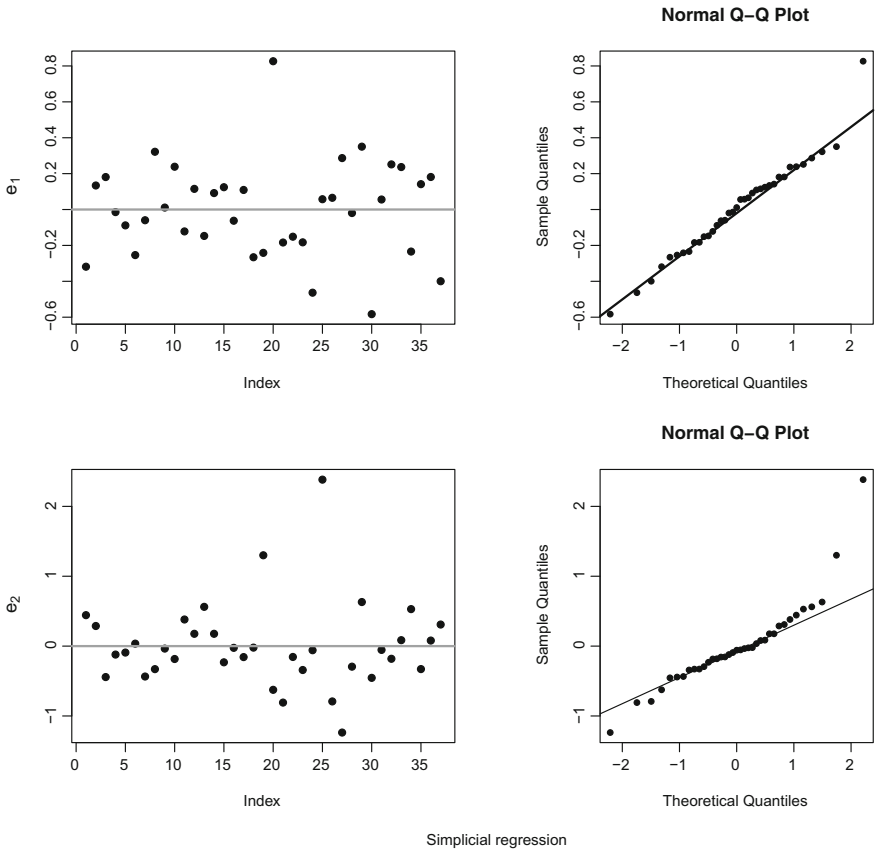

Fig. 3 In the *left column* coordinate residual plots associated to the estimated simplicial regression. In the *right column* Q–Q plots for residuals of the corresponding regression models are displayed

Moreover, the Aitchison distance of any two compositions $\mathbf{x}$ and $\mathbf{y} \in \mathscr{S}^D$ is defined as

$$d_a(\mathbf{x}, \mathbf{y}) = \sqrt{\frac{1}{2D} \sum_{i=1}^{D} \sum_{r=1}^{D} \left(\ln \frac{x_i}{x_r} - \ln \frac{y_i}{y_r} \right)^2} . \tag{20}$$

Similarly to the standard least squares regression analysis, the compositional total sum of squares (CSST) and the compositional sum of squared residuals (CSSE) are given by $CSST = \sum_{j=1}^{n} d_a^2(\mathbf{x}_j, \mathrm{g_m}(\mathbf{x}))$ and $CSSE = \sum_{j=1}^{n} d_a^2(\mathbf{x}_j, \hat{\mathbf{x}}_j)$. In this way another R^2-measure based on the compositional sum of squares [13] is

$$R_A^2 = 1 - \frac{CSSE}{CSST} . \tag{21}$$

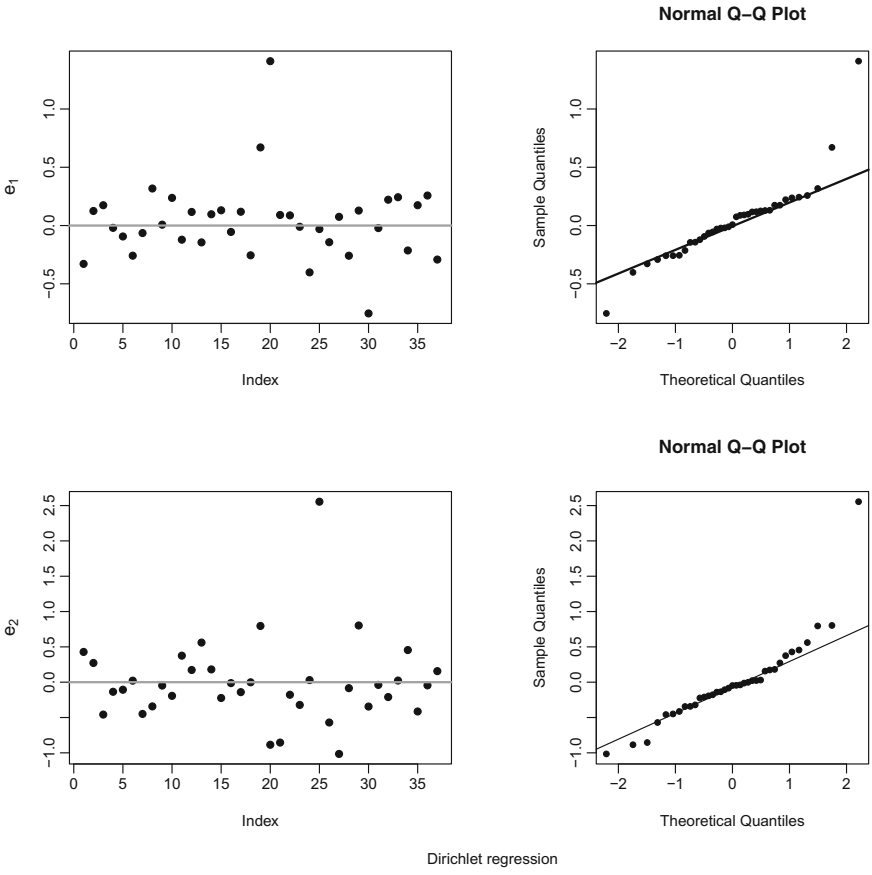

Fig. 4 In the *left column* coordinate residual plots associated to the estimated Dirichlet regression. In the *right column* Q–Q plots for residuals of the corresponding regression models are displayed

In Table 5, KL-div refers to the Kullback–Leibler divergence calculated as

$$\sum_{j=1}^{n}\sum_{i=1}^{3} x_{ji}\log\frac{x_{ji}}{\hat{x}_{ji}}\,.$$

The measures of goodness of fit reported in Table 5 show a good performance of the Shifted-Dirichlet model with respect to the other two models. The coefficient of determination based on the Aitchison norm shows that 60 % of the total variability is captured by the Shifted-Dirichlet regression model.

In order to check for absence of trends in central tendency and in variability, diagnostic plots are useful. We have expressed all the regression models in orthonormal

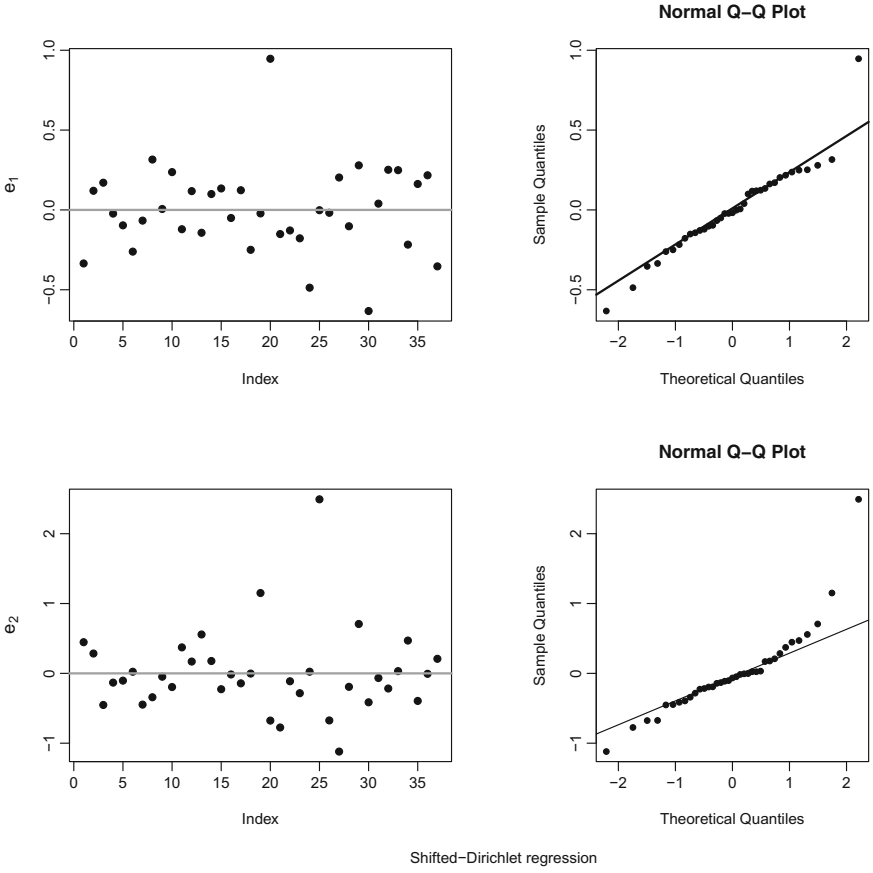

Fig. 5 In the *left column* coordinate residual plots associated to the estimated Shifted-Dirichlet regression. In the *right column* Q–Q plots for residuals of the corresponding regression models are displayed

coordinates (see Eq. (16)) which is making it possible to apply the standard battery of testing hypotheses for linear regression models, such as testing marginal normality of each coordinate residual. Coordinate residual plots and associated normal Q–Q plots for the three different regression models are displayed in Figs. 3, 4 and 5.

Except for the tails of the distribution (see Fig. 5), the assumption of normality seems to be reasonable. For the first coordinate residuals, whose points are displayed in the upper left of Fig. 5, the p-values of the Anderson–Darling test and of the Lilliefors (Kolmogorov–Smirnov) test for normality are 0.19 and 0.4789, respectively, so that the hypothesis of normal distribution cannot be rejected, while for the second coordinate residuals, lower left of Fig. 5, the two p-values of the two mentioned normality tests are 0.0009 and 0.003, respectively, due to the presence of an upper outlier (25th observation). If we omit such outlying point, normality is confirmed, i.e., the p-value of Anderson–Darling test equals 0.431 while the Lilliefors test p-value is 0.115.

5 Conclusions

Regression models with compositional response were proposed in the eighties. In this work, using the Shifted-Dirichlet distribution a new covariate model on the simplex is proposed. The Shifted-Dirichlet distribution is a generalization of the Dirichlet distribution obtained, within the Aitchison geometry, after applying a perturbation to the standard Dirichlet one. As a probability distribution, it is the same as the Scaled-Dirichlet distribution but the density is expressed with respect to the Aitchison measure on the simplex, and not with respect to the Lebesgue measure in the induced Euclidean geometry from the real space. Consequently, the Shifted-Dirichlet regression model is a generalization of the Dirichlet regression model. Even though the number of parameters to estimate increases, we see that it is a feasible and more flexible model. Using a real data set, we obtain results comparable to those obtained using the Simplicial regression.

Acknowledgments Research partially financially supported by the Italian Ministry of University and Research, FAR (Fondi di Ateneo per la Ricerca) 2012, by the Ministerio de Economía y Competividad under the projects METRICS (Ref. MTM2012-33236) and CODA-RETOS (Ref. MTM2015-65016-C2-1-R), and by the Agència de Gestió d'Ajuts Universitaris i de Recerca of the Generalitat de Catalunya under the project COSDA (Ref: 2014SGR551).

References

1. Abramowitz, M., Stegun, I.A.: Handbook of Mathematical Functions with Formulas, Graphs, and Mathematical Tables. Dover Publications, New York (1964)
2. Aitchison, J.: The Statistical Analysis of Compositional Data. Chapman & Hall Ltd. (Reprinted in 2003 with additional material by The Blackburn Press), London (1986)

3. Aitchison, J., Shen, S.M.: Logistic-normal distributions. Some properties and uses. Biometrika **67**, 261–272 (1980)
4. Billheimer, D., Guttorp, P., Fagan, W.F.: Statistical interpretation of species composition. J. Am. Stat. Assoc. **96**, 1205–1214 (2001)
5. Cameron, A.C., Windmeijer, F.A.G.: R-squared measures for count data regression models with applications to health-care utilization. J. Busin. Econ. Stat. **14**(2), 209–220 (1996)
6. Campbell, G., Mosimann, J.: Multivariate methods for proportional shape. In: ASA Proceedings of the Section on Statistical Graphics (1987)
7. Coakley, J.P., Rust, B.R.: Sedimentation in an Arctic lake. J. Sediment. Petrol. **38**, 1290–1300 (1968)
8. Egozcue, J.J., Daunis-i-Estadella, J., Pawlowsky-Glahn, V., Hron, K., Filzmoser, P.: Simplicial regression. The normal model. J. App. Prob. Stat. **6**, 87–108 (2012)
9. Egozcue, J.J., Pawlowsky-Glahn, V.: Groups of parts and their balances in compositional data analysis. Math. Geol. **37**(7), 795–828 (2005)
10. Ferrari, S., Cribari-Neto, F.: Beta regression for modelling rates and proportions. J. App. Stat. **31**, 799–815 (2004)
11. Garzanti, E., Vezzoli, G., Andó, S.: Paleogeographic and paleodrainage changes during Pleistocene glaciations (Po Plain, Northern Italy). Earth-Sci. Rev. **105**, 25–48 (2011)
12. Gueorguieva, R., Rosenheck, R., Zelterman, D.: Dirichlet component regression and its applications to psychiatric data. Comput. Stat. Data Anal. **52**, 5344–5355 (2008)
13. Hijazi, R.H.: Residuals and Diagnostics in Dirichlet regression. Tech Report of United Arab Emirates University, Department of Statistics (2006)
14. Hijazi, R.H., Jernigan, R.W.: Modelling compositional data using dirichlet regression models. J. App. Prob. Stat. **4**, 77–91 (2009)
15. Mateu-Figueras, G., Pawlowsky-Glahn, V., Egozcue, J.J.: The principle of working on coordinates. In: Compositional Data Analysis, pp. 29–42. John Wiley & Sons, Ltd (2011)
16. Mateu-Figueras, G., Pawlowsky-Glahn, V., Egozcue, J.J.: The normal distribution in some constrained sample spaces. SORT **37**(2), 231–252 (2011)
17. Melo, T.F.N., Vasconcellos, K.L.P., Lemonte, A.J.: Some restriction tests in a new class of regression models for proportions. Comput. Stat. Data Anal. **53**, 3972–3979 (2009)
18. Monti, G.S., Mateu-Figueras, G., Pawlowsky-Glahn, V.: Notes on the scaled Dirichlet distribution. In: Compositional Data Analysis, pp. 128–138. John Wiley & Sons, Ltd (2011)
19. Monti, G.S., Mateu-Figueras, G., Pawlowsky-Glahn, V., Egozcue, J.J.: The shifted-scaled Dirichlet distribution in the simplex. In: Proceedings of The 4th International Workshop on Compositional Data Analysis (2011)
20. Monti, G.S., Mateu-Figueras, G., Pawlowsky-Glahn, V. and Egozcue, J.J.: Scaled-Dirichlet covariate models for compositional data. In: Proceedings of 47th Scientific Meeting of the Italian Statistical Society (2014)
21. Pawlowsky-Glahn, V.: Statistical modelling on coordinates. In: Proceedings of Compositional Data Analysis Workshop—CoDaWork'03 (2003)
22. Pawlowsky-Glahn, V., Egozcue, J.J.: Geometric approach to statistical analysis on the simplex. Stoch Env. Res. Risk A. **15**, 384–398 (2001)
23. Pawlowsky-Glahn, V., Egozcue, J.J., Tolosana-Delgado, R.: Modelling and analysis of compositional data, p. 272. Statistics in practice. John Wiley & Sons, Chichester UK
24. R Development Core Team: R: A Language and Environment for Statistical Computing. R Foundation for Statistical Computing (2014)
25. Rayens, W.S., Srinivasan, C.: Dependence properties of generalized Liouville distributions on the simplex. J. Am. Stat. Ass. **89**, 1465–1470 (1994)