Journal of Geochemical Exploration 255 (2023) 107327

Contents lists available at ScienceDirect

Journal of Geochemical Exploration

journal homepage: www.elsevier.com/locate/gexplo

Lasso regression method for a compositional covariate regularised by the norm L^1 pairwise logratio

Jordi Saperas-Riera *, Glòria Mateu-Figueras, Josep Antoni Martín-Fernández

Universitat de Girona, C/de la Universitat de Girona, 6, Girona 17003, Spain

ARTICLE INFO

Keywords:
Aitchison's geometry
Compositional data
Norm L^1
Balance selection

ABSTRACT

Lasso regression methods include a penalty function expressed in terms of a norm defined in the space of model coefficients. The norm plays a key role as regards the way coefficients can become irrelevant in the model. For models with a compositional covariate, the norm should be coherent with the Aitchison geometry. The proposed method is based on a newly-defined compositional norm called L^1 pairwise logratio. The novel approach allows one to construct an appropriate basis through a sequential binary partition for discriminating between balances that influence the response variable and those that have no effect. This generalised Lasso regression scheme is illustrated with the analysis of a geochemical data set.

1. Introduction

One of the goals of linear regression analysis is to identify a subset of explanatory variables that are associated with the response variable. For example, in geochemistry it may be of interest to identify which chemical elements have an important effect on the soil pH in a particular region. To address this, Lasso regression methods, introduced by Tibshirani (1996), are a popular option for variable selection. Lasso regression applies an L^1-norm penalisation to the model coefficients (*slopes*), where the L^1-norm is the sum of the absolute value of the coefficients. The standard regression models assume the independence of the covariates, having each one its own slope. Importantly, these assumptions do not apply to a compositional explanatory variable, that is, in the case of compositional data (CoDa).

CoDa analysis (Aitchison, 1986) has become increasingly important in various fields such as environmental science, geochemistry, microbiology, and economics. However, CoDa poses unique challenges, especially when compositions are used as covariates in regression models. Indeed, following the *principle of working on coordinates* (Mateu-Figueras et al., 2011), the D-part composition in the explanatory part of the model should be expressed in terms of at least $D-1$ logarithms of ratios of raw variables (*logratios*). Recently, a number of papers provided tools and methods for regression model simplification with CoDa. First works on penalised regression with compositional covariates are Lin et al. (2014); Shi et al. (2016); Lu et al. (2019), later extended to robust regression in Monti and Filzmoser (2021, 2022). Some of them are focused on considering all possible pairwise logratios in a penalised regression model (Bates and Tibshirani, 2019; Susin et al., 2020; Calle and Susin, 2022a, 2022b; Calle et al., 2023) with the usual L^1-norm (Lasso) or L^2-norm (Ridge) or a linear combination of both (Elastic net) applied to the model coefficients. Other works use supervised learning methods to select pairwise logratios in a generalised linear model (Coenders and Greenacre, 2022). A pairwise logratio approach in CoDa analysis is based on comparing the logarithm of the ratios between two parts of a composition. This approach allows one to analyse the relative information between different parts while avoiding issues of scale dependence and spurious correlation (Aitchison, 1986). Importantly, an approach based on balances can be considered as a generalisation because a balance is a logarithm of the ratios between the average of two groups of parts (Egozcue and Pawlowsky-Glahn, 2005). Balances in CoDa analysis are useful for identifying geochemical relationships and gaining insights into geological processes. By examining the ratios of different elements within samples, researchers can determine patterns and potential causes of variation. This approach can be applied to a range of materials, from rocks and minerals to soils and sediments, and can inform our understanding of issues (Buccianti and Grunsky, 2014). In the context of linear regression models for CoDa, Rivera-Pinto et al. (2018) propose a stepwise algorithm for selecting balances but the global optimum is not guaranteed. A more efficient algorithm identifying a sequence of balances is introduced by Gordon-Rodriguez et al. (2022). In addition, Nesrstová et al. (2023) introduce a Partial Least Squares procedure to construct principal balances (Martín-Fernández

* Corresponding author.
E-mail addresses: jordi.saperas@udg.edu (J. Saperas-Riera), gloria.mateu@udg.edu (G. Mateu-Figueras), josepantoni.martin@udg.edu (J.A. Martín-Fernández).

https://doi.org/10.1016/j.gexplo.2023.107327
Received 6 August 2023; Received in revised form 21 September 2023; Accepted 29 September 2023
Available online 18 October 2023

et al., 2018) that maximise the explained variability of the response variable.

To our knowledge, none of these recent works deals with identifying a particular structure of the parts in a composition for selecting a subset of parts (*subcomposition*) in a linear regression model. On the other hand, Boogaart et al. (2021) deal with compositional part selection by introducing the concepts of *internal and external subcompositional independence*. In a linear regression model, a subcomposition is internal independent when changes in values within the parts of the subcomposition have no effect on the explained variable. Whereas external independence further assumes that the balance between the parts of the subcomposition and the rest of the parts in the composition also does not influence the response variable. Once one detects a subcomposition both internally and externally independent then the subset of parts can be removed from the model. The key element of the method proposed in this article is the new norm called L^1 pairwise logratio (L^1-plr) taking part in the penalty term of the Lasso regression model. Using this norm, the Lasso method is able to automatically identify internal independent subcompositions, that is, it detects which pairwise logratios are influential in the response variable and which are not. Once the independent subcompositions have been identified, the model can be simplified by considering an adequate set of balances involving the corresponding parts. Furthermore, a Bootstrap scheme is proposed for checking the external independence and, in such a case, indicating the parts that can be removed from the model.

This article is organised as follows. In Section 2 the basic concepts of CoDa and standard penalty regression are described. In Section 3 the new norm L^1-plr is introduced, and its compositional properties are provided. In Section 4 the Lasso regression model using the norm L^1-plr is formulated and its properties are explored. A geochemical case study is provided in Section 5 for illustration purposes. Finally, the last section concludes with some remarks.

The analyses discussed in this article were carried out in R (R-Core-Team, 2022) using the packages *ADMM* (You and Zhu, 2021) and *coda.base* (Comas-Cufí, 2022).

2. Some basic concepts

2.1. Compositional data

CoDa conveys relative information because the variables describe relative contributions to a given total (Aitchison, 1986). These variables are called *parts* of a whole and are, usually, expressed in proportions, percentages or ppm. Historically (Aitchison, 1986), the sample space of CoDa is designed as the D-part unit simplex $\mathscr{S}^D = \{\mathbf{x} \in \mathbb{R}^D : x_j > 0; \sum x_j = 1; j = 1,\ldots,D\}$. The formal geometric framework for the analysis of CoDa first appeared in Pawlowsky-Glahn and Egozcue (2001) and Billheimer et al. (2001). This geometry was coined the *Aitchison geometry*, later formally established in Barceló-Vidal and Martn-Fernández (2016). The property of scale invariance of results in the analysis offers a broader understanding of compositions. According to this property, two vectors one multiple of the other are considered compositionally equivalent. Consequently, the set of vectors proportional to $\mathbf{x} \in \mathscr{S}^D$ ($\{k\cdot\mathbf{x};\ k > 0\}$) is called a composition and for simplicity denoted again by $\mathbf{x}$. While the compositional space is the set of all compositions and it is denoted for simplicity by $\mathscr{S}^D$.

The Aitchison geometry is based on two specific operations that induce a vector space structure on $\mathscr{S}^D$ called *perturbation* and *powering*, and defined as $\mathbf{x} \oplus \mathbf{y} = (x_1y_1, x_2y_2, \ldots, x_Dy_D)$ and $\alpha \odot \mathbf{x} = \left(x_1^\alpha, x_2^\alpha, \ldots, x_D^\alpha\right)$ for $\mathbf{x},\mathbf{y} \in \mathscr{S}^D$, $\alpha \in \mathbb{R}$. In order to interpret the results of these operations, one can *closure* the result, that is, to normalise the resulting vector to a unit sum by dividing each component by its total sum. Note that the closure operation provides a vector compositionally equivalent.

Once we have a vector space structure, a metric structure is easily defined using the clr-scores of a composition $\mathbf{x}$ (Aitchison, 1986):

$$\mathrm{clr}(\mathbf{x}) = \left(\mathrm{clr}(\mathbf{x})_1, \ldots, \mathrm{clr}(\mathbf{x})_D\right) = \left(ln\frac{x_1}{g(\mathbf{x})}, \ldots, ln\frac{x_D}{g(\mathbf{x})}\right),$$

where $g(\cdot)$ is the geometric mean of the composition. Indeed, the basic metric elements of the Aitchison geometry: inner product ($<\cdot,\cdot>_{\mathscr{A}}$), L^2-norm ($\|\cdot\|_{\mathscr{A}}$), and distance ($d_{\mathscr{A}}(\cdot,\cdot)$) are

$$<\mathbf{x},\mathbf{y}>_{\mathscr{A}} = <\mathrm{clr}(\mathbf{x}), \mathrm{clr}(\mathbf{y})>_E \quad , \quad \|\mathbf{x}\|^2_{\mathscr{A}} = <\mathbf{x},\mathbf{x}>_{\mathscr{A}} \quad , \quad d_{\mathscr{A}}(\mathbf{x},\mathbf{y}) = \|\mathbf{x} \ominus \mathbf{y}\|_{\mathscr{A}}, \tag{1}$$

where "$\mathscr{A}$" means the Aitchison geometry, "E" means the typical Euclidean geometry, and "$\ominus$" is the perturbation difference $\mathbf{x} \ominus \mathbf{y} = \mathbf{x} \oplus ((-1) \odot \mathbf{y})$.

An important scale invariant function is the *logcontrast* because it plays the typical role of the linear combination of variables. Given a composition $\mathbf{x} = (x_1, \ldots, x_D)$, a logcontrast is defined as any linear combination of the logarithms of the compositional parts:

$$\sum_{j=1}^{D} a_j ln x_j, \quad \text{with} \quad \sum_{j=1}^{D} a_j = 0, \quad a_j \in \mathbb{R}. \tag{2}$$

Note that each clr-score $\mathrm{clr}(\mathbf{x})_j; j = 1, \ldots, D$, is a logcontrast, and, on the other side, any logcontrast can be expressed as a logratio

$$\sum_{j=1}^{D} a_j ln x_j = ln\frac{\prod_{a_j>0} x_j^{a_j}}{\prod_{a_j<0} x_j^{|a_j|}}.$$

In fact, parts with $a_j > 0$ in (2) appear in the numerator, and parts with $a_j < 0$ appear in the denominator. If a part has no contribution, then $a_j = 0$.

The metric elements defined in Eq. (1) can be used to construct an orthonormal logratio (*olr*) basis and to calculate the corresponding olr-coordinates of a composition ($\mathrm{olr}(\mathbf{x})$, formerly known as ilr-coordinates) (Egozcue et al., 2003; Martín-Fernández, 2019). The expression of these olr-coordinates depends on the basis selected. Following Egozcue and Pawlowsky-Glahn (2005), one can define particular olr-coordinates, called *balances*. A balance involves two groups of parts of a composition and is expressed as the logratio of the geometric mean of each group of parts multiplied by a constant to guarantee the unit length of the vectors of the basis.

A sequential binary partition (SBP) of a composition $\mathbf{x} = (x_1, \ldots, x_D)$ provides balances associated with a specific olr-basis. In the first step of an SBP, the full composition $\mathbf{x} = (x_1, \ldots, x_D)$ is split into two groups of parts: one for the numerator (coded with $+1$) and the other for the denominator (with code -1). According to this partition, the first olr-coordinate is obtained as the logarithm of the geometric mean of the parts in the numerator divided by the geometric mean of the parts in the denominator, multiplied by a scaling factor that depends on the number of parts (Eq. (3)). In the following steps, each group of parts is in turn split into two groups and the following olr-coordinates are obtained. In step k when the $\mathrm{olr}(\mathbf{x})_k$-coordinate is created, the r_k parts $(x_{n1_k}, \ldots, x_{nr_k})$ in the first group are placed in the numerator (code $+1$); the s_k parts $(x_{d1_k}, \ldots, x_{ds_k})$ in the second group will appear in the denominator (code -1); and the rest of $D - (r_k + s_k)$ parts are not involved in the logratio (code 0). As a result, the $\mathrm{olr}(\mathbf{x})_k$ is:

$$\mathrm{olr}(\mathbf{x})_k = \sqrt{\frac{r_k \cdot s_k}{r_k + s_k}} ln \frac{\left(x_{n1_k} \cdots x_{nr_k}\right)^{1/r_k}}{\left(x_{d1} \cdots x_{ds_k}\right)^{1/s_k}}, \quad k = 1, \ldots, D-1, \tag{3}$$

where $\sqrt{\frac{r_k \cdot s_k}{r_k + s_k}}$ is the factor for normalising vectors of the basis. Note that the $\mathrm{olr}(\mathbf{x})_k$ coordinate, being a logcontrast that involves two groups of parts, informs us of, on average, the relative importance of one group of parts with regard to the other.

Relating the clr-scores with any olr-coordinates by means of a matrix relationship is straightforward. Indeed, $\mathrm{olr}_\Psi(\mathbf{x}) = \Psi \mathrm{clr}(\mathbf{x})$ and $\mathrm{clr}(\mathbf{x}) =$

$\Psi^T \text{olr}_\Psi(\mathbf{x})$, with $\Psi \in \mathbb{R}^{(D-1)\times D}$ a matrix where the $D-1$ rows are the clr-scores of compositions forming the olr-basis. Consequently, all compositional operations and compositional metric elements (Eq. (1)) are translated into ordinary operations between the corresponding olr-coordinates.

2.2. Linear model with compositional covariates

Given a dependent variable y and an explanatory D-part composition $\mathbf{x}$, the definition of a linear regression model in terms of a logcontrast (Aitchison and Bacon-Shone, 1984; Hron et al., 2012) is:

$$y = \alpha_0 + \sum_{j=1}^{D} \alpha_j ln x_j, \quad \text{with} \quad \sum_{j=1}^{D} \alpha_j = 0, \quad \alpha_j \in \mathbb{R}, \tag{4}$$

whereas, in terms of metric elements, the model formulation is (Boogaart and Tolosana, 2013):

$$y = \beta_0 + \langle \boldsymbol{\beta},\ \mathbf{x} \rangle_{\mathscr{A}} = \beta_0 + \langle \text{clr}(\boldsymbol{\beta}),\ \text{clr}(\mathbf{x}) \rangle_E = \beta_0 + \langle \text{olr}_\Psi(\boldsymbol{\beta}),\ \text{olr}_\Psi(\mathbf{x}) \rangle_E, \tag{5}$$

where $\boldsymbol{b}$ is the compositional gradient vector. Note that $\Psi \in \mathbb{R}^{(D-1)\times D}$ is the matrix associated to any olr-basis in the clr-space, for example, a basis created using an SBP. Considering the expression in terms of clr-scores, the coefficients could be estimated using a statistical toolbox but the use of the generalised inversion for the covariance matrix of the clr-scores is required (Boogaart et al., 2021), which it is not necessary when working with the model based on olr-coordinates.

2.3. Penalty regression

The Lasso regression model is formulated as the combination of the L^2-norm cost function and the L^1-norm regularisation term. For a real data set $\mathbf{X}$ with n observations and D predictors and a real response vector $\mathbf{Y}$ of length n, the Lasso regression model can be formulated as (Tibshirani, 1996)

$$min \left\{ \frac{1}{2} \| \mathbf{Y} - \beta_0 - \mathbf{X}\cdot\boldsymbol{\beta} \|_2^2 + \lambda \| \boldsymbol{\beta} \|_1 \right\}, \tag{6}$$

where a is the intercept, $\mathbf{b}$ is the gradient, and λ is the penalty parameter that controls the amount of regularisation. For $\lambda = 0$, the Lasso regression model (Eq. (6)) provides the classical least squares regression model. The larger the value of λ, the greater the number of coefficients in $\mathbf{b}$ forced to be zero. The *optimal* value of λ can be chosen based on cross-validation techniques or other methods (James et al., 2021).

It is possible to generalise the Lasso problem by taking the L^1-norm of a linear transformation of gradient $\mathbf{b}$ as a penalty function. The generalised Lasso regression model is

$$min \left\{ \frac{1}{2} \| \mathbf{Y} - \beta_0 - \mathbf{X}\cdot\boldsymbol{\beta} \|_2^2 + \lambda \| \mathbf{F}\cdot\boldsymbol{\beta} \|_1 \right\}, \tag{7}$$

where $\mathbf{F}$ is the matrix associated to an arbitrary linear transformation. Note that $\mathbf{F} = Id$ corresponds to the simple Lasso problem.

The metric elements (Eq. (1)) used to define the regression model (Eq. (5)) facilitate the formulation of the cost function in a Lasso model Eqs. (6) and (7) for compositional covariates. However, the definition of an appropriate L^1-norm for CoDa requires the supplementary concepts described in the following section.

3. Norm L^1 pairwise logratio

The Aitchison norm $\| \mathbf{x} \|_{\mathscr{A}}$ (Eq. (1)) is defined as the Euclidean L^2-norm of the clr-scores ($L^2 - \text{clr}$):

$$\| \mathbf{x} \|_{\mathscr{A}}^2 = \| \mathbf{x} \|_{2-\text{clr}}^2 = \| \text{clr}(\mathbf{x}) \|_2^2 = \sum_{j=1}^{D} \left(ln\left(\frac{x_j}{g(\mathbf{x})} \right) \right)^2, \tag{8}$$

that is, $\| \mathbf{x} \|_{\mathscr{A}}$ can be interpreted as the restriction of a L^2 Euclidean norm on the clr-space. Following this idea, the norm $L^1 - \text{clr}$ ($\| \mathbf{x} \|_{1-\text{clr}}$) for a Lasso regression model can be defined as (Bates and Tibshirani, 2019; Susin et al., 2020)

$$\| \mathbf{x} \|_{1-\text{clr}} = \| \text{clr}(\mathbf{x}) \|_1 = \sum_{j=1}^{D} \left| ln\left(\frac{x_j}{g(\mathbf{x})} \right) \right|. \tag{9}$$

Note that a regularisation term $\| \boldsymbol{\beta} \|_{1-\text{clr}}$ forces some components $\text{clr}(\boldsymbol{\beta})_j, \quad j = 1, \ldots, D$, to take small values, suggesting the corresponding parts $\mathbf{x}_j$ could be removed from the model. However, the presence of the removed parts in the geometrical mean of the non-removed parts ($g(\boldsymbol{\beta})$ and $g(\mathbf{x})$) is a difficulty for the regression model simplification.

Importantly, the Aitchison distance (Eq. (1)) can also be defined in terms of pairwise logratios (Aitchison et al., 2000). Consequently, the Aitchison norm can be expressed as $\| \mathbf{x} \|_{\mathscr{A}}^2 = \frac{1}{D} \sum_{i<j} \left(ln\left(\frac{x_i}{x_j} \right) \right)^2$. Based on this expression, a new L^1-norm on the simplex $\mathscr{S}^D$ is defined as:

Definition 1. The L^1-plr norm of a composition $\mathbf{x} \in \mathscr{S}^D$ is

$$\| \mathbf{x} \|_{1-plr} = \frac{1}{D-1} \sum_{i<j} \left| ln\left(\frac{x_i}{x_j} \right) \right|. \tag{10}$$

Proposition 2. $\| \mathbf{x} \|_{1-plr} = \frac{1}{D-1} \sum_{i<j} \left| ln\left(\frac{x_i}{x_j} \right) \right|$ *verifies the properties of a norm. That is:*

- *Positive definiteness:* $\forall \mathbf{x} \in \mathscr{S}^D, \ \| \mathbf{x} \|_{1-plr} \geq 0$. *Moreover,* $\| \mathbf{x} \|_{1-plr} = 0$ *if and only if* $\mathbf{x} = (1, \ldots, 1)$.
- *Absolute homogeneity:* $\forall \mathbf{x} \in \mathscr{S}^D$ *and* $\forall \lambda \in \mathbb{R}, \quad \| \lambda \odot \mathbf{x} \|_{1-plr} = |\lambda| \, \| \mathbf{x} \|_{1-plr}$.
- *Subadditivity:* $\forall \mathbf{x}, \ \mathbf{y} \in \mathscr{S}^D, \ \| \mathbf{x} \oplus \mathbf{y} \|_{1-plr} \leq \| \mathbf{x} \|_{1-plr} + \| \mathbf{y} \|_{1-plr}$.

See Appendix for the proof.

Importantly, the coefficient accompanying the sum of squared pairwise logratios in the Aitchison norm is $1/D$, whereas in the norm L^1-plr is $1/(D-1)$ (Eq. (10)). Using this factor, the norm L^1-plr is endowed with the property of subcompositional dominance among other compositional properties:

Proposition 3. *The* L^1*-plr norm on* $\mathscr{S}^D$, $\| \mathbf{x} \|_{1-plr} = \frac{1}{D-1} \sum_{i<j} \left| ln\left(\frac{x_i}{x_j} \right) \right|$ *verifies the properties*

- *Scale invariance:* $\| \mathbf{x} \|_{1-plr} = \| \lambda \mathbf{x} \|_{1-plr}, \ \lambda > 0$.
- *Permutation invariance:* $\| (x_1, \ldots, x_i, \ldots, x_j, \ldots, x_D) \|_{1-plr} = \| (x_1, \ldots, x_j, \ldots, x_i, \ldots, x_D) \|_{1-plr}$.
- *Subcompositional dominance:* $\| \mathbf{x} \|_{1-plr} \geq \| \text{sub}(\mathbf{x}) \|_{1-plr}$ *where* $\text{sub}(\mathbf{x})$ *denotes any subcomposition of* $\mathbf{x}$.

See Appendix for the proof.

The norm L^1-plr can be expressed in terms of the clr-scores as

$$\| \mathbf{x} \|_{1-plr} = \frac{1}{D-1} \sum_{i<j} \left| ln\left(\frac{x_i / g(\mathbf{x})}{x_j / g(\mathbf{x})} \right) \right| = \frac{1}{D-1} \sum_{i<j} \left| \text{clr}(\mathbf{x})_i - \text{clr}(\mathbf{x})_j \right|, \tag{11}$$

suggesting that, in general, $\| \mathbf{x} \|_{1-plr} \neq \| \mathbf{x} \|_{1-clr}$.

Fig. 1 shows the shape of the unit balls (i.e., set of points that have distance 1 from the origin) measured by the norms L^1-clr (Eq. (9), green), L^1-plr (Eq. (10), orange), and Aitchison (Eq. (1), blue) in the 3-part compositional space. To represent it, the olr-coordinates $\text{olr}_1(\mathbf{x}) = \frac{\sqrt{2}}{2} ln \frac{x_1}{x_2}$, and $\text{olr}_2(\mathbf{x}) = \sqrt{\frac{2}{3}} ln \frac{\sqrt{x_1 x_2}}{x_3}$ are used. Because the norms are calculated using clr-scores and pairwise logratios, the value of the norms is

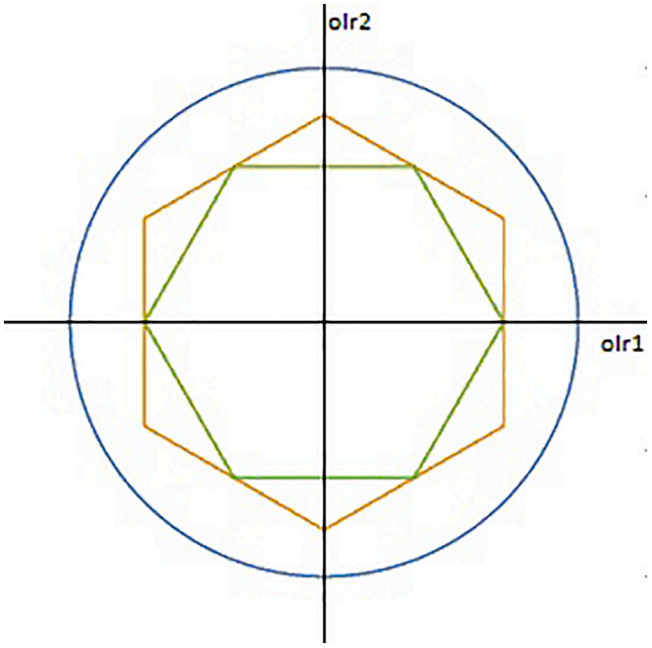

Fig. 1. Unit balls in the olr-coordinates space of 3-part compositions using the norms: L^2-clr (blue), L^1-clr (green), and L^1-plr (orange). (For interpretation of the references to colour in this figure legend, the reader is referred to the web version of this article.)

invariant under a change of olr-basis. When one takes a different olr-basis then the shape of the unit balls is only affected by a rotation. As expected, the unit ball of the L^2 Aitchison norm (blue) shows the typical shape of a circle, which includes the L^1 unit balls as it is well known for norms L^p in Euclidean spaces. Interestingly, both norms L^1-clr (green) and L^1-plr (orange) create a hexagon, latter being the biggest. That is, the unit ball with norm L^1-plr includes the unit ball of norm L^1-clr because it holds $\| \mathbf{x}\|_{1-plr} \leq \| \mathbf{x}\|_{1-clr}$, for $\mathbf{x} \in \mathscr{S}^D$ (see Appendix for the proof). The points of contact between the unit balls correspond to $\mathbf{x} \in \mathscr{S}^3$ that $\mathrm{clr}(\mathbf{x}) \in \left\{ \left(\pm\frac{1}{2}, \mp\frac{1}{2}, 0\right); \left(\pm\frac{1}{2}, 0, \mp\frac{1}{2}\right); \left(0, \pm\frac{1}{2}, \mp\frac{1}{2}\right) \right\}$, where $\| \mathbf{x}\|_{1-\mathrm{plr}} = \| \mathbf{x}\|_{1-\mathrm{clr}} = 1$.

4. Generalised Lasso regression with the norm L^1 pairwise logratio

To our knowledge, the compositional Lasso regression methods introduced in the literature (Bates and Tibshirani, 2019; Susin et al., 2020; Calle and Susin, 2022a, 2022b; Calle et al., 2023) aim to separate the parts into two groups: parts that influence the response variable, and parts that do not. However, the methods do not analyse the *external independence* of parts that do not affect the response variable (Boogaart et al., 2021). That is, they do not explore the balance between the non-influential subcomposition and the rest of the parts.

We will show that the Lasso regression using our new norm L^1-plr aims to identify and separate the balances (i.e., pairwise logratios) into two groups: the balances that influence the response variable, and those that do not. Therefore this method permits the analyst to deal with both types of subcompositional independence: *internal* and *external* (Boogaart et al., 2021).

Definition 4. Given y_i, $i = 1, \ldots, n$ the sample of the response variable, $\mathbf{X}$ the $n \times D$ matrix whose rows, $\mathbf{X}_i = (x_{i1}, \ldots, x_{iD})$ for $i = 1, \ldots, n$, contains the compositional sample, and $\mathrm{clr}(\mathbf{X})_i$ the i-th row of matrix $\mathrm{clr}(\mathbf{X})$. The L^1-plr Lasso estimator is defined as

$$\boldsymbol{\beta} \in \underset{\boldsymbol{\beta}}{argmin} \left\{ \frac{1}{2} \sum_{i=1}^{n} \left(y_i - \beta_0 - \langle \mathrm{clr}(\boldsymbol{\beta}),\ \mathrm{clr}(\mathbf{X})_i \rangle_E \right)^2 + \lambda \| \boldsymbol{\beta}\|_{1-plr} \right\}. \tag{12}$$

Following Eq. (11), it holds that $\| \boldsymbol{\beta}\|_{1-plr} = \| \mathbf{F} \cdot \mathrm{clr}(\boldsymbol{\beta})\|_1$, where $\mathbf{F}$ is the matrix $\frac{D(D-1)}{2} \times D$ associated to the linear transformation $F(z_1, \ldots, z_D) = \frac{1}{D-1}(z_1 - z_2, z_1 - z_3, \ldots, z_1 - z_D, z_2 - z_3, \ldots, z_2 - z_D, \ldots, z_{D-1} - z_D)$. Consequently, Definition 4 can be generalised to:

$$\boldsymbol{\beta} \in \underset{\boldsymbol{\beta}}{argmin} \left\{ \frac{1}{2} \sum_{i=1}^{n} \left(y_i - \beta_0 - \langle \mathrm{clr}(\boldsymbol{\beta}),\ \mathrm{clr}(\mathbf{X})_i \rangle_E \right)^2 + \lambda \| \mathbf{F} \cdot \mathrm{clr}(\boldsymbol{\beta})\|_1 \right\}. \tag{13}$$

The matrix $\mathrm{clr}(\mathbf{X})$ is not a full rank matrix, thus causing troubles when solving the convex optimisation problem (Saperas-Riera et al. (2023)) in Eq. (13). To avoid these troubles the problem can be solved in terms of olr_Ψ-coordinates, $\mathrm{olr}_\Psi(\mathbf{x}) = \Psi \cdot \mathrm{clr}(\mathbf{x})$, that is, the L^1-plr Lasso estimator (Eq. (12)) in olr_Ψ-coordinates is

$$\boldsymbol{\beta} \in \underset{\boldsymbol{\beta}}{argmin} \left\{ \frac{1}{2} \sum_{i=1}^{n} \left(y_i - \beta_0 - \langle \mathrm{olr}_\Psi(\boldsymbol{\beta}),\ \mathrm{olr}_\Psi(\mathbf{X})_i \rangle_E \right)^2 + \lambda \| \mathbf{F} \cdot \Psi^T \cdot \mathrm{olr}_\Psi(\boldsymbol{\beta})\|_1 \right\}. \tag{14}$$

The relationship between Eq. (12) and Eq. (14) can be used for analyzing the relations between coefficients $\mathrm{clr}(\boldsymbol{\beta})_j$; $j = 1, \ldots, D$ and the coefficients of the balances in the generalised Lasso model (Eq. (14)). Importantly, the penalty term in Eq. (13) forces the sum of the absolute value of the differences of the clr-scores of the gradient vector to be less than a fixed number, which forces some pairwise differences of clr-scores to be zero ($\mathrm{clr}(\boldsymbol{\beta})_i - \mathrm{clr}(\boldsymbol{\beta})_j = 0$), that is, forces some pairs of clr-scores to be equal. This means that the corresponding pairwise logratios ($ln\frac{x_i}{x_j}$) do not influence the response variable. For example, without loss of generality, suppose that the pairwise logratio $ln\frac{x_1}{x_2}$ does not influence the response variable y. That is, the coefficient of the balance in the model is equal to zero. In this case, taking an adequate matrix Ψ one can obtain a model in clr-scores with $\mathrm{clr}(\boldsymbol{\beta})_1 = \mathrm{clr}(\boldsymbol{\beta})_2$ (i.e., $\beta_1 = \beta_2$). And vice-versa, if $\boldsymbol{\beta}$ fulfils $\beta_1 = \beta_2$, an olr-basis including the unit vector $\frac{\sqrt{2}}{2}(1, -1, 0, \ldots, 0)$ provides a model where the coefficient of the pairwise logratio $ln\frac{x_1}{x_2}$ is equal to zero, that is, the pairwise logratio $ln\frac{x_1}{x_2}$ does not influence the response variable y. The above reasoning can be extended to a linear model with gradient $\boldsymbol{\beta}$ fulfilling $\beta_1 = \beta_2 = \ldots = \beta_k$, $2 < k < D$. In this case, with an adequate matrix Ψ, one detects that the any pairwise logratio $ln(x_i/x_j)$, $1 \leq i < j \leq k$ do not influence the response variable y. In addition, any balance involving some of the parts in the subcomposition $(x_1, \ldots, x_k)$ does not influence the response variable y. That is, the subcomposition is *internal independent* (Boogaart et al., 2021).

Following the idea described above, a general algorithm for a L^1-plr Lasso method can be formulated as:

Algorithm 1. L^1-plr Lasso.

1. Fit the L^1-plr Lasso model with tuning parameter λ (Eq. (14)).
2. Express the L^1-plr Lasso model in terms of clr-scores (Eq. (5)).
3. For each string detected being $\left\{ \mathrm{clr}(\boldsymbol{\beta})_{j_1} = \ldots = \mathrm{clr}(\boldsymbol{\beta})_{j_k} \right\}$, built an orthonormal basis for the subcomposition $(\mathbf{x}_{j_1}, \ldots, \mathbf{x}_{j_k})$ in $\mathscr{S}^D$ (Eq. (3)).
4. Put together the bases created above for the subcompositions. Complete until an olr-basis basis for the full composition $\mathbf{x} \in \mathscr{S}^D$ is reached. Write the L^1-plr Lasso model in terms of the olr-coordinates (Eq. (5)).

When fitting the model in $\mathrm{olr} - \textit{coordinates}$ (step 1), any matrix Ψ can be used. Once fitted, the relationship between clr-scores and olr-coordinates ($\mathrm{clr}(\mathbf{x}) = \Psi^T \mathrm{olr}_\Psi(\mathbf{x})$) is used for detecting the subcompositions of the gradient vector $\boldsymbol{\beta}$ fulfiling $\beta_{j_1} = \ldots = \beta_{j_k}$, $2 \leq k < D$ (step 2). The

balances forming the olr-bases of these subcompositions do not influence the variable response y (step 3). That is, in this step, internal independent subcompositions are detected. The rest of the parts of the composition $\mathbf{x} \in \mathscr{S}^D$ form an influential subcomposition. When completing the olr-basis for the full composition $\mathbf{x}$, any olr-basis can be created for the influential subcomposition (step 4). Importantly, the corresponding balances linking the different subcompositions detected must be included in the olr-basis. Moreover, the significance of the coefficient of a linking balance between a non-influential subcomposition and the rest of the parts provides information about the *external independence*. Consequently, any subcomposition $\{\mathbf{x}_{j_1}, \ldots, \mathbf{x}_{j_k}\}$, $2 \leq k < D$ both internal and external independent can be removed from the linear regression model. A routine written in R code (R-Core-Team, 2022) has been developed by us to perform the steps involved in carrying out the algorithm. The routine is freely available from the leading author.

5. Case study

Following Boogaart et al. (2021), a total of $n = 2095$ samples of the data set of project GEMAS ("Geochemical Mapping of Agricultural and grazing land Soil") were analyzed. For further information about the data set, you can consult Reimann et al. (2014a, 2014b). The analyzed data set contains information on the *soil pH*, as a real response variable y. The compositional covariate is the 11-part composition $\mathbf{x}$ of the major oxides and LOI (loss on ignition): (SiO_2, TiO_2, Al_2O_3, Fe_2O_3, MnO, MgO, CaO, Na_2O, K_2O, P_2O_5, LOI).

To select the optimal λ parameter for our model in Eq. (14), a 10-fold cross-validation was performed. Each iteration involves dividing the data into 10 equal parts, training the model on nine of them, and then evaluating it on the remaining part to produce the lowest Mean Squared Error (MSE). Fig. 2, shows a plot with the MSE curve and the values of $lambda.min = 4.197$ and the $lambda.1se = 52.436$. With the value $lambda.min$, one obtains the minimum mean cross-validated error, whereas $lambda.1se$ is the largest value of the tuning parameter λ such that the error is within one standard error of the cross-validated errors for $lambda.min$. Because the larger value of an optimal λ, the larger the number of regularised coefficients, the value $lambda.1se = 52.436$ was considered for the L^1 Lasso model (James et al., 2021).

The linear penalised model for $\lambda = lambda.1se = 52.436$ expressed in terms of clr-scores (Eq. (5)) has the following coefficients: intercept $\beta_0 = 5.938$ and gradient

$$\mathrm{clr}(\boldsymbol{\beta}) = (0.046, 0.046, 0.046, 0.046, 0.096, 0.083, 0.550, -0.489, 0.096, -0.182, -0.339).$$

In this case, the method detects two strings: $\beta_1 = \beta_2 = \beta_3 = \beta_4$ and $\beta_5 = \beta_9$. Consequently, the associated balances and/or pairwise logratios within the subcompositions $(SiO_2, TiO_2, Al_2O_3, Fe_2O_3)$ and (MnO, K_2O) do not have any influence on the response variable *soil pH*. The first 4-part subcomposition forms a 3-dimensional logratio subspace where all the pairwise logratios and balances involving the major oxides SiO_2, TiO_2, Al_2O_3, and Fe_2O_3 are non-influential. For example, the pairwise logratio $ln\frac{SiO_2}{Al_2O_3}$ or the balance $ln\frac{(SiO_2Al_2O_3)^{1/2}}{(TiO_2Fe_2O_3)^{1/2}}$ are non-influential. The second subcomposition defines a 1-dimensional space, that is, changes in the values of the pairwise logratio $ln\frac{MnO}{K_2O}$ have no effect on the soil pH. On the other hand, pairwise logratios and balances mixing major oxides of the first group $(SiO_2, TiO_2, Al_2O_3, Fe_2O_3)$ with major oxides of the second group (MnO, K_2O) could be influential. In particular, the coefficient for the linking balance between both subcompositions could be significant. The linking balance is the second

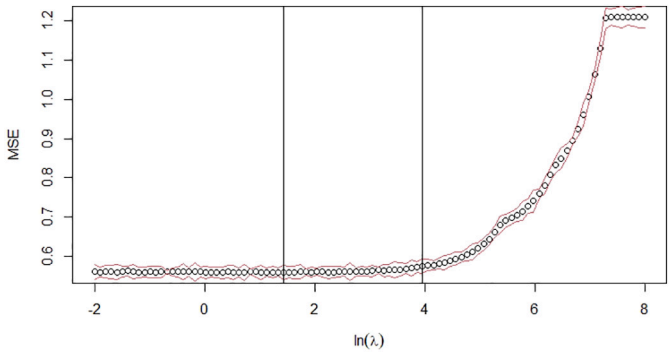

Fig. 2. Cross-validation MSE curve for different log-transformed values of the penalty parameter ($ln(\lambda)$). The circle (∘) is the arithmetical mean of the 10-fold CV. The red lines (above and below the mean) are respectively the value $mean \pm stdev$, where $stdev$ is the standard deviation of the 10-fold CV. Vertical lines are the log-transformed values of $lambda.min = 4.197$ and $lambda.1se = 52.436$. (For interpretation of the references to colour in this figure legend, the reader is referred to the web version of this article.)

balance (olr_2, blue) in Table 1. The other balances that can be influential are those involving the rest of parts forming the subcomposition (MgO, CaO, Na_2O, P_2O_5, LOI) and the linking balance between the two types of subcompositions (Table 1, green).

Table 1 summarises the subcompositional structure suggested by the L^1-plr Lasso method when one creates an appropriate olr-basis using an SBP. Note that the code "+ 1"in the SBP means that the part is in the numerator of the balance, whereas for the parts in the denominator, the label is "− 1". The code "0"is reserved for the parts non-involved in the balance. In green, the first row is the full balance between the major oxides involved in the non-influential subcompositions and the remaining major oxides ($MgO, CaO, Na_2O, P_2O_5, LOI$). In blue, olr_2 balances the two non-influential subcompositions. In red, the three rows form the basis of the subcomposition (SiO_2, TiO_2, Al_2O_3, Fe_2O_3). In purple, the sixth row is the vector forming basis of the subcomposition (MnO, K_2O). In black, the last rows show an SBP for creating an olr-basis of the subcomposition with the remaining major oxides.

The L^1-plr Lasso gradient $\boldsymbol{\beta}$ expressed in terms of olr-coordinates associated to the SBP defined in Table 1 is

$$\mathrm{olr}_\Psi(\boldsymbol{\beta}) = (0.228, -0.058, 0, 0, 0, 0, 0.194, 0.735, 0.236, 0.187),$$

where coefficients equal to zero correspond to non-influential balances. Accordingly, the linear regression model is

$$y = 5.938 + 0.228\ \mathrm{olr}_1(\mathbf{x}) - 0.058\ \mathrm{olr}_2(\mathbf{x}) + 0.194\ \mathrm{olr}_7(\mathbf{x}) + 0.735\ \mathrm{olr}_8(\mathbf{x}) + 0.236\ \mathrm{olr}_9(\mathbf{x}) + 0.187\ \mathrm{olr}_{10}(\mathbf{x}). \quad (15)$$

Note that the largest coefficient is $\mathrm{olr}_\Psi(\boldsymbol{\beta})_8 = 0.7345$, suggesting that the ratio $\frac{CaO}{Na_2O}$ concentrates most of the predicting power for pH. That is, one can interpret that when increasing the ratio $\frac{CaO}{Na_2O}$ while keeping all other predictor balances constant the soil pH increases (Coenders and Pawlowsky-Glahn, 2020). This feature coincides with the evaluation of the model presented in Boogaart et al. (2021). Among the other coefficients, it is of particular interest to test if the coefficient of the balance $\mathrm{olr}_2(\mathbf{x})$ is zero ($\mathrm{olr}_\Psi(\boldsymbol{\beta})_2 = -0.058$). Because olr_2 is the linking balance between the two non-influential subcompositions (Table 1), removing this balance would indicate a non-influential 6-part

Table 1
SBP and balances for the olr-basis suggested by the L^1-plr Lasso method. Colours are associated with subcompositions (see text for details).

$D = 11$											
SiO_2	TiO_2	Al_2O_3	Fe_2O_3	MnO	MgO	CaO	Na_2O	K_2O	P_2O_5	LOI	*Balances*
+1	+1	+1	+1	+1	−1	−1	−1	+1	−1	−1	$\mathrm{olr}_1 = \sqrt{\frac{30}{11}} \ln \frac{(SiO_2\ TiO_2\ Al_2O_3\ Fe_2O_3\ MnO\ K_2O)^{1/6}}{(MgO\ CaO\ Na_2O\ P_2O_5\ LOI)^{1/5}}$
+1	+1	+1	+1	−1	0	0	0	−1	0	0	$\mathrm{olr}_2 = \sqrt{\frac{4}{3}} \ln \frac{(SiO_2\ TiO_2\ Al_2O_3\ Fe_2O_3)^{1/4}}{(MnO\ K_2O)^{1/2}}$
+1	+1	−1	−1	0	0	0	0	0	0	0	$\mathrm{olr}_3 = \ln \frac{(SiO_2\ TiO_2)^{1/2}}{(Al_2O_3\ Fe_2O_3)^{1/2}}$
+1	−1	0	0	0	0	0	0	0	0	0	$\mathrm{olr}_4 = \frac{\sqrt{2}}{2} \ln \frac{SiO_2}{TiO_2}$
0	0	+1	−1	0	0	0	0	0	0	0	$\mathrm{olr}_5 = \frac{\sqrt{2}}{2} \ln \frac{Al_2O_3}{Fe_2O_3}$
0	0	0	0	+1	0	0	0	−1	0	0	$\mathrm{olr}_6 = \frac{\sqrt{2}}{2} \ln \frac{MnO}{K_2O}$
0	0	0	0	0	−1	+1	+1	0	−1	−1	$\mathrm{olr}_7 = \sqrt{\frac{6}{5}} \ln \frac{(CaO\ Na_2O)^{1/2}}{(MgO\ P_2O_5\ LOI)^{1/3}}$
0	0	0	0	0	0	+1	−1	0	0	0	$\mathrm{olr}_8 = \frac{\sqrt{2}}{2} \ln \frac{CaO}{Na_2O}$
0	0	0	0	0	+1	0	0	0	+1	−1	$\mathrm{olr}_9 = \sqrt{\frac{2}{3}} \ln \frac{(MgO\ P_2O_5)^{1/2}}{LOI}$
0	0	0	0	0	+1	0	0	0	−1	0	$\mathrm{olr}_{10} = \frac{\sqrt{2}}{2} \ln \frac{MgO}{P_2O_5}$

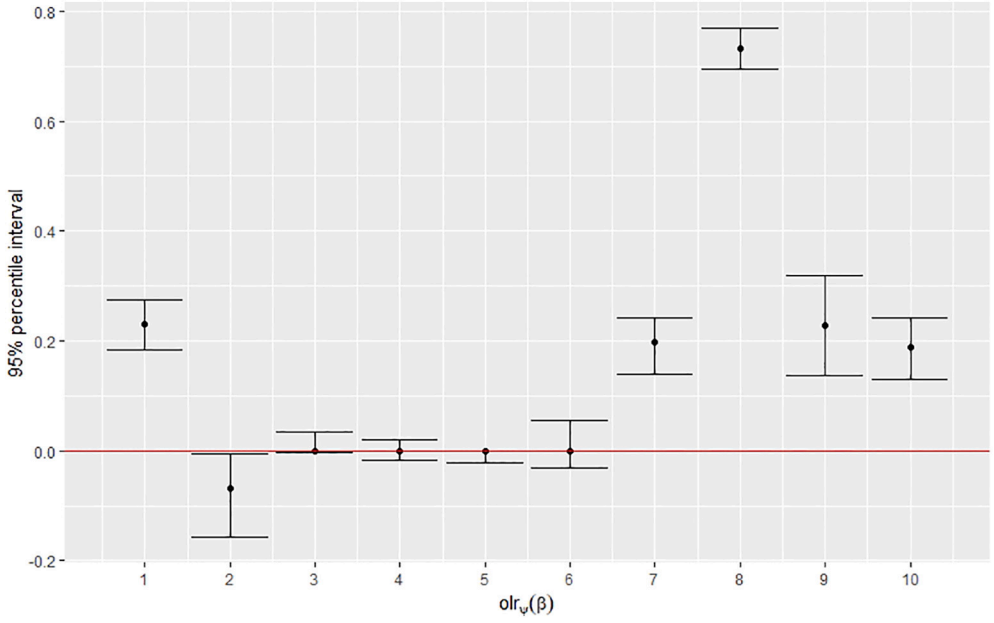

Fig. 3. Bootstrapping 95 % percentile intervals for the coefficients $\mathrm{olr}_\Psi(\boldsymbol{\beta})$ of the linear regression model.

subcomposition $(SiO_2,\ TiO_2,\ Al_2O_3,\ Fe_2O_3,\ MnO,\ K_2O)$, generating a 5-dimensional space. In addition, eliminating the balance $\mathrm{olr}_1(\mathbf{x})$ would simplify the model because the subcompositional external independence of the subcomposition would permit removing major oxides SiO_2, TiO_2, Al_2O_3, Fe_2O_3, MnO, and K_2O from the model. In this case, the coefficient involved is $\mathrm{olr}_\Psi(\boldsymbol{\beta})_1 = 0.228$.

Following Hesterberg et al. (2012), one can add the uncertainty associated with the coefficient using a bootstrap technique. Fixed $\lambda = 52.436$ and the SBP (Table 1), 1.000 random samplings with replacement were performed in the data set and the L^1-plr Lasso model was fitted. As a result, the 95 % percentile interval for each coefficient in $\mathrm{olr}_\Psi(\beta)$ was calculated. Fig. 3 shows that the coefficient $\mathrm{olr}_\Psi(\boldsymbol{\beta})_1 = 0.228$ cannot be considered equal to zero, indicating that the subcomposition $(SiO_2,\ TiO_2,\ Al_2O_3,\ Fe_2O_3,\ MnO,\ K_2O)$ cannot be removed from the model because it is not external independent. In addition, because the 95 % percentile interval for the coefficient $\mathrm{olr}_\Psi(\beta)_2 \in (-0.158, -0.005)$ does not contain the zero then one can assume that the balance $\mathrm{olr}_2(\mathbf{x})$ is influential, i.e., it cannot be removed from the model. Finally, as expected, percentile intervals of coefficients $\mathrm{olr}_\Psi(\beta)_3$, $\mathrm{olr}_\Psi(\beta)_4$, $\mathrm{olr}_\Psi(\beta)_5$, and $\mathrm{olr}_\Psi(\beta)_6$ contain the zero because the internal independence of subcomposition $(SiO_2,\ TiO_2,\ Al_2O_3,\ Fe_2O_3)$ and $(MnO,\ K_2O)$. Consequently, the model in Eq. (15) does not admit further simplification.

Despite L^1-plr Lasso removing four balances from the model (Eq. (15)), all the major oxides participate in the remaining balances. That is, no parts have been removed from the model. On the other hand, when one applies the L^1-clr Lasso method (Lin et al., 2014; Bates and Tibshirani, 2019) the goal is selecting influential parts. In the case of the GEMAS data set, the optimal λ tuning parameter was selected ($lambda.1se = 38.728$) using a 10-fold cross-validation evaluating the MSE. The linear regression model created has intercept $\beta_0 = 6.440$ and

the clr-gradient:

$$\mathrm{clr}(\boldsymbol{\beta}) = (0,\ 0,\ 0,\ 0,\ 0.162, 0.081, 0.555, -0.489, 0.170, -0.142, -0.336).$$

That is, the 4-part subcomposition of major oxides $(SiO_2, TiO_2, Al_2O_3, Fe_2O_3)$ has been removed from the linear regression model:

$$y = 6.44 + 0.162lnMnO + 0.081lnMgO + 0.555lnCaO - 0.489lnNa_2O + 0.170lnK_2O - 0.142lnP_2O_5 - 0.336lnLOI.$$

Note that the coefficients of this model have a lack of interpretation (Coenders and Pawlowsky-Glahn, 2020). For example, despite the coefficient of CaO being 0.555, it is not possible to interpret that one should expect an increase in soil pH when CaO is increased while the other concentrations are kept constant. Because concentrations are relative data (CoDa), the proportion of one part cannot increase if the other proportions are kept. Consequently, an adequate interpretation of the model requires being expressed in terms of balances or pairwise logratios (Coenders and Pawlowsky-Glahn, 2020).

6. Final remarks

Because concentrations of geochemical elements are compositional, any statistical analysis has to consider their relative nature. This article fills the gap for methods for automatic recognition of internal independent subcompositions in linear regression models. We introduced a new norm for CoDa, the norm L^1 pairwise logratio. This norm verifies the desirable properties, such as scale invariance and subcompositional dominance, in order to be coherent with the Aitchison geometry. By using the norm L^1 pairwise logratio in a Lasso regression model, we can determine the importance of the relative information between the compositional parts in explaining a response variable. We use this information to create an olr-basis taking into account the structure defined by the internal independent subcompositions. The basis created includes linking balances between the internal independent subcompositions and the rest of the parts of the composition. We test the linking balances for analysing the subcompositional external independence because in such a case, the parts involved can be removed from the regression model. In other words, because the method identifies the pairwise logratios and balances that are less relevant, we can simplify the model and improve its interpretation while maintaining a high level of predictive power. This methodology provides a more nuanced understanding of the relationship between the compositional parts, which can lead to better insights and decision-making in various fields. Still pending is analyzing how one can improve the model when introducing a penalty term based on a convex linear combination of the norm L^1-plr and the Aitchison norm (Elastic net). The development of these types of models is one of the more interesting challenges in current CoDa analysis. Moreover, in order to implement the algorithm on high-dimensional data sets is necessary to increase the speed when fitting the model. One option to explore could be to change the ADMM algorithm by a faster one.

Declaration of competing interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Data availability

The authors do not have permission to share data.

Acknowledgments

This research was supported by the Ministerio de Ciencia e Innovación under the project "CODA-GENERA" (Ref. PID2021-123833OB-I00) and the grant PRE2019-090976; and by the Agència de Gestió d'Ajuts Universitaris i de Recerca of the Generalitat de Catalunya under the project "COSDA" (Ref. 2021SGR01197).

Appendix A. Proofs

Proposition. 2. $\| \mathbf{x}\|_{1-plr} = \frac{1}{D-1}\sum_{i<j}|ln\left(\frac{x_i}{x_j}\right)|$ *verifies the properties of a norm.*

Proof.
- Positive definiteness: $\forall \mathbf{x} \in \mathscr{S}^D,\ \| \mathbf{x}\|_{1-plr} \geq 0$. Moreover, $\| \mathbf{x}\|_{1-plr} = 0$ if and only if $\mathbf{x} = (1, \ldots, 1)$. Immediate from definition 1.
- Absolute homogeneity: $\forall \mathbf{x} \in \mathscr{S}^D$ and $\forall \lambda \in \mathbb{R},\ \| \lambda \odot \mathbf{x}\|_{1-plr} = |\lambda| \| \mathbf{x}\|_{1-plr}$. Immediate from definition 1.
- Subadditivity: $\forall \mathbf{x},\ \mathbf{y} \in \mathscr{S}^D,\ \| \mathbf{x} \oplus \mathbf{y}\|_{1-plr} \leq \| \mathbf{x}\|_{1-plr} + \| \mathbf{y}\|_{1-plr}$.

Because the absolute value is a convex function, for all $i,\ j$, it verifies $\left|ln\left(\frac{x_iy_i}{x_jy_j}\right)\right| = \left|ln\left(\frac{x_i}{x_j}\right) + ln\left(\frac{y_i}{y_j}\right)\right| \leq \left|ln\left(\frac{x_i}{x_j}\right)\right| + \left|ln\left(\frac{y_i}{y_j}\right)\right|$. Thus, $\frac{1}{D-1}\sum_{i<j}\left|ln\left(\frac{x_iy_i}{x_jy_j}\right)\right| \leq \frac{1}{D-1}\sum_{i<j}\left|ln\left(\frac{x_i}{x_j}\right)\right| + \frac{1}{D-1}\sum_{i<j}\left|ln\left(\frac{y_i}{y_j}\right)\right|$. □

Proposition. 3. *The L^1-plr norm on $\mathscr{S}^D$, $\| \mathbf{x}\|_{1-plr} = \frac{1}{D-1}\sum_{i<j}|ln\left(\frac{x_i}{x_j}\right)|$ verifies the properties scale invariance, permutation invariance, and subcompositional dominance.*

Proof. .

- Scale invariance: $\| \mathbf{x}\|_{1-plr} = \| \lambda\mathbf{x}\|_{1-plr}$. Immediate from definition 1.
- Permutation invariance: $\| (x_1, \ldots, x_i, \ldots, x_j, \ldots, x_D)\|_{1-plr} = \| (x_1, \ldots, x_j, \ldots, x_i, \ldots, x_D)\|_{1-plr}$. Immediate from definition 1.
- Subcompositional dominance: $\| \mathbf{x}\|_{1-plr} \geq \| \mathrm{sub}(\mathbf{x})\|_{1-plr}$.

Without loss of generality we will prove that $\| (x_1, \ldots, x_D)\|_{1-plr} \geq \| (x_2, \ldots, x_D)\|_{1-plr}$.

$$\| \mathbf{x}\|_{1-plr} = \frac{1}{D-1}\sum_{i<j}\left|ln\left(\frac{x_i}{x_j}\right)\right| = \frac{1}{D-1}\left(\sum_{j=2}^{D}\left|ln\left(\frac{x_1}{x_j}\right)\right| + \sum_{2\leq i<j}\left|ln\left(\frac{x_i}{x_j}\right)\right|\right).$$

We write the first summation, $\sum_{j=2}^{D}\left|ln\left(\frac{x_1}{x_j}\right)\right|$, in a double summation form:

$$\sum_{j=2}^{D}\left|ln\left(\frac{x_1}{x_j}\right)\right| = \frac{1}{D-2}\sum_{j=2}^{D}(D-2)\left|ln\left(\frac{x_1}{x_j}\right)\right| = \frac{1}{D-2}\sum_{j=2}^{D}\sum_{k=3}^{D}\left|ln\left(\frac{x_1}{x_k}\right)\right| =$$

$$\frac{1}{D-2}\sum_{2\le j<k}\left(\left|ln\left(\frac{x_1}{x_j}\right)\right| + \left|ln\left(\frac{x_1}{x_k}\right)\right|\right)$$

Renaming index j by i, and index k by j in the above summation, we can write the L^1-plr norm as follows:

$$\| \mathbf{x}\|_{1-plr} = \frac{1}{D-1}\sum_{2\le i<j}\left(\frac{1}{D-2}\left|ln\left(\frac{x_1}{x_i}\right)\right| + \frac{1}{D-2}\left|ln\left(\frac{x_1}{x_j}\right)\right| + \left|ln\left(\frac{x_i}{x_j}\right)\right|\right) \ge$$

$$\frac{1}{D-1}\sum_{2\le i<j}\left(\frac{1}{D-2}\left|ln\left(\frac{x_i}{x_j}\right)\right| + \left|ln\left(\frac{x_i}{x_j}\right)\right|\right) = \frac{1}{D-1}\sum_{2\le i<j}\frac{D-1}{D-2}\left|ln\left(\frac{x_i}{x_j}\right)\right|.$$

Therefore,

$$\| \mathbf{x}\|_{1-plr} = \frac{1}{D-1}\sum_{i<j}\left|ln\left(\frac{x_i}{x_j}\right)\right| \ge \frac{1}{D-1}\frac{D-1}{D-2}\sum_{2\le i<j}\left|ln\left(\frac{x_i}{x_j}\right)\right| = \| (x_2,\ldots,x_D)\|_{1-plr}.$$

Note the importance of using the factor $\frac{1}{D-1}$ in the norm L^1-plr instead of the factor $\frac{1}{D}$ used in the Aitchison norm. □

Proposition. . *For all* $\mathbf{x}\in\mathscr{S}^D$, *it holds that* $\| \mathbf{x}\|_{1-\mathrm{plr}} \le \| \mathbf{x}\|_{1-\mathrm{clr}}$.

Proof. .

$$\| \mathbf{x}\|_{1-plr} = \frac{1}{D-1}\sum_{i<j}\left|ln\left(\frac{x_i}{x_j}\right)\right| = \frac{1}{D-1}\sum_{i<j}\left|ln\left(\frac{x_i/g(\mathbf{x})}{x_j/g(\mathbf{x})}\right)\right| = \frac{1}{D-1}\sum_{i<j}\left|ln\left(\frac{x_i}{g(\mathbf{x})}\right) - ln\left(\frac{x_j}{g(\mathbf{x})}\right)\right| \le$$

$$\frac{1}{D-1}\sum_{i<j}\left|ln\left(\frac{x_i}{g(\mathbf{x})}\right)\right| + \left|ln\left(\frac{x_j}{g(\mathbf{x})}\right)\right|$$

For each $k=1,\ldots,D$, the term $ln\left(\frac{x_k}{g(\mathbf{x})}\right)$ appears $D-1$ times in the last summation, then it holds that

$$\frac{1}{D-1}\sum_{i<j}\left|ln\left(\frac{x_i}{g(\mathbf{x})}\right)\right| + \left|ln\left(\frac{x_j}{g(\mathbf{x})}\right)\right| = \| \mathbf{x}\|_{1-\mathrm{clr}}.$$

Therefore, it holds that $\mathbf{x}\in\mathscr{S}^D$, $\| \mathbf{x}\|_{1-\mathrm{plr}} \le \| \mathbf{x}\|_{1-\mathrm{clr}}$.

Note that for any $\mathbf{x}\in\mathscr{S}^D$ that $\mathrm{clr}(\mathbf{x}) = \left(0,\ldots,0,\underbrace{\pm\frac{a}{2}}_{i},0,\ldots,0,\underbrace{\mp\frac{a}{2}}_{j},0,\ldots,0\right)$ then it holds that $\| \mathbf{x}\|_{1-\mathrm{plr}} = \| \mathbf{x}\|_{1-\mathrm{clr}} = a$. □

References

Aitchison, J., 1986. The Statistical Analysis of Compositional Data. Chapman & Hall, London. Reprinted 2003 with additional material by The Blackburn Press, London, UK.

Aitchison, J., Bacon-Shone, J., 1984. Log contrast models for experiments with mixtures. Biometrika 71, 323–330.

Aitchison, J., Barceló-Vidal, C., Martín-Fernández, J., Pawlowsky-Glahn, V., 2000. Logratio analysis and compositional distance. Math. Geol. 32, 271–275.

Barceló-Vidal, C., Martn-Fernández, J.A., 2016. The mathematics of compositional analysis. Austrian J. Stat. 45, 57–71.

Bates, S., Tibshirani, R., 2019. Log-ratio lasso: scalable, sparse estimation for log-ratio models. Biometrics 75, 613–624.

Billheimer, D., Guttorp, P., Fagan, W.F., 2001. Statistical interpretation of species composition. J. Am. Stat. Assoc. 96, 1205–1214. https://doi.org/10.1198/016214501753381850 arXiv:https://doi.org/10.1198/016214501753381850.

Boogaart, K.G.v.d., Tolosana, R., 2013. Analyzing Compositional Data with R. Use R! Springer.

Boogaart, K., Filzmoser, P., Hron, K., Templ, M., Tolosana-Delgado, R., 2021. Classical and robust regression analysis with compositional data. Math. Geosci. 53, 823–858.

Buccianti, A., Grunsky, E., 2014. Compositional data analysis in geochemistry: are we sure to see what really occurs during natural processes? J. Geochem. Explor. 141 https://doi.org/10.1016/j.gexplo.2014.03.022.

Calle, M., Susin, A., 2022a. coda4microbiome: compositional data analysis for microbiome studies. bioRxiv doi:https://doi.org/10.1101/2022.06.09.495511.

Calle, M., Susin, A., 2022b. Identification of dynamic microbial signatures in longitudinal studies. bioRxiv doi:https://doi.org/10.1101/2022.04.25.489415.

Calle, M., Pujolassos, M., Susin, A., 2023. coda4microbiome: compositional data analysis for microbiome cross-sectional and longitudinal studies. BMC Bioinform. https://doi.org/10.1186/s12859-023-05205-3.

Coenders, G., Greenacre, M., 2022. Three approaches to supervised learning for compositional data with pairwise logratios. J. Appl. Stat. https://doi.org/10.1080/02664763.2022.2108007.

Coenders, G., Pawlowsky-Glahn, V., 2020. On interpretations of tests and effect sizes in regression models with a compositional predictor. SORT-Stat. Oper. Res. Trans. 44, 201–220. URL: https://raco.cat/index.php/SORT/article/view/371189 https://doi.org/10.2436/20.8080.02.100.

Comas-Cufí, M., 2022. coda.base: A Basic Set of Functions for Compositional Data Analysis. URL: https://CRAN.R-project.org/package=coda.base. r package version 0.5.2.

Egozcue, J.J., Pawlowsky-Glahn, V., 2005. Groups of parts and their balances in compositional data analysis. Math. Geol. 37, 795–828.

Egozcue, J.J., Pawlowsky-Glahn, V., Mateu-Figueras, G., Barceló-Vidal, C., 2003. Isometric logratio transformations for compositional data analysis. Math. Geol. 35, 279–300.

Gordon-Rodriguez, E., Quinn, T.P., Cunningham, J.P., 2022. Learning sparse log-ratios for high-throughput sequencing data. Bioinformatics 38, 157–163.

Hesterberg, T., Moore, D.S., Monaghan, S., Clipson, A., Epstein, R., Craig, B.A., McCabe, G., 2012. Bootstrap Methods and Permutation Tests, 7th ed. W. H. Freeman, New York, p. 657. Chapter 16 of Introduction to the Practice of Statistics.

Hron, K., Filzmoser, P., Thompson, K., 2012. Linear regression with compositional explanatory variables. J. Appl. Stat. 39, 1–14. https://doi.org/10.1080/02664763.2011.644268.

James, G., Witten, D., Hastie, T., Tibshirani, R., 2021. Introduction to Statistical Learning, 2nd edition. Springer, New York.

Lin, W., Shi, R., Feng, R., Li, H., 2014. Variable selection in regression with compositional covariates. Biometrika 101, 785–797.

Lu, J., Shi, P., Li, H., 2019. Generalized linear models with linear constraints for microbiome compositional data. Biometrics 75, 235–244.

Martín-Fernández, J.A., 2019. Comments on: compositional data: the sample space and its structure. TEST 28, 653–657.

Martín-Fernández, J.A., Pawlowsky-Glahn, V., Egozcue, J.J., Tolosona-Delgado, R., 2018. Advances in principal balances for compositional data. Math. Geosci. 50, 273–298.

Mateu-Figueras, G., Pawlowsky-Glahn, V., Egozcue, J.J., 2011. The Principle of Working on Coordinates. Chapter 3 of Compositional Data Analysis: Theory and Applications, pp. 29–42. https://doi.org/10.1002/9781119976462.ch3.

Monti, G., Filzmoser, P., 2021. Sparse least trimmed squares regression with compositional covariates for high-dimensional data. Bioinformatics 37, 3805–3814.
Monti, G., Filzmoser, P., 2022. Robust logistic zero-sum regression for microbiome compositional data. ADAC 16, 301–324.
Nesrstová, V., Wilms, I., Palarea-Albaladejo, J., Filzmoser, P., Martín-Fernández, J., Friedecký, D., Hron, K., 2023. Principal balances of compositional data for regression and classification using partial least squares. J. Chemom., e3518 https://doi.org/10.1002/cem.3518.
Pawlowsky-Glahn, V., Egozcue, J.J., 2001. Geometric approach to statistical analysis on the simplex. Stoch. Env. Res. Risk A. 15, 384–398. https://doi.org/10.1007/s004770100077.
R-Core-Team, 2022. R: A Language and Environment for Statistical Computing. R Foundation for Statistical Computing, Vienna, Austria.
Reimann, C., Birke, M., Demetriades, A., Filzmoser, P., O'Connor, P. (Eds.), 2014a. Chemistry of Europe's Agricultural Soils—Part A: Methodology and Interpretation of the GEMAS Data Set, Geologisches Jahrbuch (Reihe B 102). Schweizerbarth, Hannover.
Reimann, C., Birke, M., Demetriades, A., Filzmoser, P., O'Connor, P. (Eds.), 2014b. Chemistry of Europe's Agricultural Soils—Part B: General Background Information and Further Analysis of the GEMAS Data Set, Geologisches Jahrbuch (Reihe B 103). Schweizerbarth, Hannover.
Rivera-Pinto, J., Egozcue, J.J., Pawlowsky-Glahn, V., Paredes, R., Noguera-Julian, M., Calle, M.L., 2018. Balances: a new perspective for microbiome analysis. MSystems 3, e00053–18.
Saperas-Riera, J., Martín-Fernández, J., Mateu-Figueras, G., 2023. Fundamentals of convex optimization for compositional data. SORT-Stat. Oper. Res. Trans. 47.
Shi, P., Zhang, A., Li, H., 2016. Regression analysis for microbiome compositional data. Ann. Appl. Stat. 10, 1019–1040.
Susin, A., Wang, Y., Lê Cao, K.A., Calle, M.L., 2020. Variable selection in microbiome compositional data analysis. NAR Genom. Bioinform. 2, lqaa029.
Tibshirani, R., 1996. Regression shrinkage and selection via the lasso. J. R. Stat. Soc. Ser. B 58, 267–288.
You, K., Zhu, X., 2021. ADMM: Algorithms Using Alternating Direction Method of Multipliers. URL: https://CRAN.R-project.org/package=ADMM. r package version 0.3.3.