mathematics

MDPI

Article

L^p-Norm for Compositional Data: Exploring the CoDa L^1-Norm in Penalised Regression

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

Department of Computer Science, Applied Mathematics and Statistics, University of Girona, 17003 Girona, Spain; jordi.saperas@udg.edu (J.S.-R.); gloria.mateu@udg.edu (G.M.-F.)
* Correspondence: josepantoni.martin@udg.edu

Abstract: The Least Absolute Shrinkage and Selection Operator (LASSO) regression technique has proven to be a valuable tool for fitting and reducing linear models. The trend of applying LASSO to compositional data is growing, thereby expanding its applicability to diverse scientific domains. This paper aims to contribute to this evolving landscape by undertaking a comprehensive exploration of the L^1-norm for the penalty term of a LASSO regression in a compositional context. This implies first introducing a rigorous definition of the compositional L^p-norm, as the particular geometric structure of the compositional sample space needs to be taken into account. The focus is subsequently extended to a meticulous data-driven analysis of the dimension reduction effects on linear models, providing valuable insights into the interplay between penalty term norms and model performance. An analysis of a microbial dataset illustrates the proposed approach.

Keywords: Aitchison's geometry; compositional data; L^p-norm; balance selection

MSC: 62J07; 62P10; 62H99

Citation: Saperas-Riera, J.; Mateu-Figueras, G.; Martín-Fernández, J.A. L^p-Norm for Compositional Data: Exploring the CoDa L^1-Norm in Penalised Regression. *Mathematics* **2024**, *12*, 1388. https://doi.org/10.3390/math12091388

Academic Editors: Antonio Di Crescenzo and Francisco Chiclana

Received: 29 January 2024
Revised: 5 April 2024
Accepted: 29 April 2024
Published: 1 May 2024

1. Introduction

Linear regression serves as a powerful framework for modelling relationships between variables, as it aims to capture the underlying patterns that govern the variability in the response variable. For instance, in the microbiome domain, there is a particular interest in identifying which taxa are associated with a variable of interest, for example, the inflammatory parameter sCD14. To address such complex problems, adopting LASSO regression methods [1] has emerged as a popular choice for variable selection. LASSO regression strategically applies the Euclidean L^1-norm penalisation to the model coefficients, wherein the L^1-norm represents the sum of the absolute values of these coefficients. The penalised term shrinks some regression parameters towards zero, facilitating variable selection.

While conventional regression models assume independence among covariates, this assumption fails when dealing with compositional explanatory variables. These variables are called *parts* of a whole, and are usually expressed in proportions, percentages, or ppm. Historically [2], the sample space of the compositional data (CoDa) is designed as the D-part unit simplex $\mathcal{S}^D = \{\mathbf{x} \in \mathbb{R}^D : x_j > 0; \sum x_j = 1; j = 1, \ldots, D\}$. The fundamental idea in the analysis of CoDa is that the information is relative, and is primarily contained in the ratios between parts, not the absolute amounts of the parts. Therefore, the use of log-ratios is advocated. The analysis of CoDa, pioneered by [2], has witnessed increasing significance across such diverse fields as environmental science, geochemistry, microbiology, and economics. However, the integration of CoDa as covariates in regression models introduces particular challenges. The existing literature addresses these challenges, providing methodologies for regression model simplification with CoDa. The first works on penalised regression with compositional covariates [3–6] restricted the Euclidean L^1-norm on the centered log-ratio (clr) subspace when defining the penalty term. Saperas et al. [7]

introduced a new norm, called the pairwise log-ratio (L^1-plr), as a part of a methodology on penalised regression to simplify the log-ratios on the explanatory side of the model. These log-ratios are also known as balances [8].

The primary objective of this article lies in comprehensive comparison of the effects of different norms on the penalty term within LASSO regression with different compositional explanatory variables. The choice of a norm in the penalty term is a pivotal aspect that significantly influences the regularization mechanism, and consequently the characteristics of the resulting models. To accomplish this, a precise and rigorous definition of the induced L^p-norms for CoDa (CoDa L^p-norms) within the compositional space is necessary. A comparison between the CoDa L^1-norm and other norms for compositions established in the literature is provided. Through this analysis, we seek to contribute valuable insights into the characteristics and implications of these norms in penalised regression.

The rest of this article is organised as follows. In Section 2, fundamental concepts related to the geometric structure of CoDa are outlined. In addition, some popular measures of central tendency are written as the solution of a variational problem using L^p-norms in real space. Section 3 is devoted to defining the CoDa L^p-norms on the compositional space. In Section 4, after describing the basic concepts of standard penalty regression, we analyse LASSO regression with compositional covariates using three different L^1-norms in the penalty term, with the CoDa L^1-norm among them. A comparison of the different norms is illustrated in Section 5 using a microbiome dataset. Finally, Section 6 concludes with some closing remarks.

2. Some Basic Concepts

2.1. Elements of the Aitchison Geometry

CoDa conveys relative information because the variables describe relative contributions to a given total [2]. The formal geometrical framework for the analysis of CoDa, coined the *Aitchison geometry*, first appeared in [9,10]. The Aitchison geometry is based on two specific operations on $\mathcal{S}^D$, called *perturbation* and *powering*, respectively defined as $\mathbf{x} \oplus \mathbf{y} = (x_1y_1, x_2y_2, \ldots, x_Dy_D)$ and $\alpha \odot \mathbf{x} = (x_1^\alpha, x_2^\alpha, \ldots, x_D^\alpha)$ for $\mathbf{x}, \mathbf{y} \in \mathcal{S}^D$, $\alpha \in \mathbb{R}$. In order to interpret the results of these operations, one can perform *closure* on the result, that is, normalise the resulting vector to a unit sum by dividing each component by its total sum. Note that the closure operation provides a compositionally equivalent vector. With a vector space structure, a metric structure can be easily defined using the clr-scores of a D-part composition $\mathbf{x} = (x_1, \ldots, x_D)$ [2]:

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

where $g(\cdot)$ is the geometric mean of the composition. Note that clr-scores are collinear, because it holds that $\sum_{j=1}^{D} \mathrm{clr}(\mathbf{x})_j = 0$.

The basic metric elements of the Aitchison geometry are the inner product ($< \cdot, \cdot >_{\mathcal{A}}$), L^2-norm ($|| \cdot ||_{\mathcal{A}}$), and distance ($d_{\mathcal{A}}(\cdot, \cdot)$), defined as follows:

$$< \mathbf{x}, \mathbf{y} >_{\mathcal{A}} = < \mathrm{clr}(\mathbf{x}), \mathrm{clr}(\mathbf{y}) >_E , \quad \|\mathbf{x}\|_{\mathcal{A}}^2 = < \mathbf{x}, \mathbf{x} >_{\mathcal{A}} , \quad d_{\mathcal{A}}(\mathbf{x}, \mathbf{y}) = \|\mathbf{x} \ominus \mathbf{y}\|_{\mathcal{A}}, \text{ for } \mathbf{x}, \mathbf{y} \in \mathcal{S}^D, \tag{1}$$

where "$\mathcal{A}$" means the Aitchison geometry, "E" the typical Euclidean geometry, and "$\ominus$" the perturbation difference $\mathbf{x} \ominus \mathbf{y} = \mathbf{x} \oplus ((-1) \odot \mathbf{y})$. Log-ratios, like clr-scores, have become a cornerstone of CoDa analysis; nevertheless, in the literature the concept of balance between two non-overlapping groups of parts is frequently used. A balance is defined as the log-ratio between the geometric means of the parts within each group multiplied by a constant that depends on the number of parts in each group [8].

An important scale-invariant function is the *log-contrast*, which plays the typical role of the linear combination of variables. Given a D-part composition $\mathbf{x}$, a log-contrast is defined as any linear combination of the logarithms of the compositional parts

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

Given a dependent variable y and an explanatory D-part composition $\mathbf{x}$, the definition of a linear regression model in terms of a log-contrast [11] is

$$y = \alpha_0 + \sum_{j=1}^{D} \alpha_j \ln x_j, \quad \text{with} \quad \sum_{j=1}^{D} \alpha_j = 0, \quad \alpha_j \in \mathbb{R}, \tag{2}$$

whereas in terms of metric elements the model formulation [12] is

$$y = \beta_0 + \langle \boldsymbol{\beta}, \mathbf{x} \rangle_{\mathcal{A}} = \beta_0 + \langle \operatorname{clr}(\boldsymbol{\beta}), \operatorname{clr}(\mathbf{x}) \rangle_E, \tag{3}$$

where $\boldsymbol{\beta}$ is the compositional gradient vector. The expressions in both Equations (2) and (3) are equivalent when one considers $\alpha_0 = \beta_0$ and $\boldsymbol{\alpha} = \operatorname{clr}(\boldsymbol{\beta})$. Because the sum of the clr-scores is zero ($\sum_{j=1}^{D} \operatorname{clr}(\beta)_j = \sum_{j=1}^{D} \operatorname{clr}(\mathbf{x})_j = 0$), the inner product of clr transformed vectors is equal to

$$\langle \operatorname{clr}(\boldsymbol{\beta}), \operatorname{clr}(\mathbf{x}) \rangle_E = \langle \operatorname{clr}(\boldsymbol{\beta}), \ln(\mathbf{x}) \rangle_E = \langle \ln \boldsymbol{\beta}, \operatorname{clr}(\mathbf{x}) \rangle_E. \tag{4}$$

For simplicity and to avoid overloading the notation, we denote $\boldsymbol{\beta}^* = \ln \boldsymbol{\beta}$, and write the linear regression model in terms of the Euclidean inner product as $y = \beta_0 + \langle \boldsymbol{\beta}^*, \operatorname{clr}(\mathbf{x}) \rangle_E$.

2.2. Norms and Measures of Central Tendency

The most popular measures of the central tendency of a real variable are the median and the arithmetic mean. Both can be defined as solving a variational problem [13]; indeed, the median $Med(\mathbf{z})$ of a dataset $\mathbf{z} = \{z_1, \ldots, z_D | \ z_i \in \mathbb{R}\}$ is the value that minimises the average absolute deviation $Med(\mathbf{z}) = \arg\min_{\lambda} \frac{1}{D} \sum_{j=1}^{D} |z_j - \lambda|$. The arithmetic mean $\overline{\mathbf{z}}$ of a dataset $\mathbf{z} = \{z_1, \ldots, z_D | \ z_i \in \mathbb{R}\}$ is the value that minimises the mean squared deviation $\overline{\mathbf{z}} = \arg\min_{\lambda} \frac{1}{D} \sum_{j=1}^{D} (z_j - \lambda)^2$. In addition, the mid-range $MR(\mathbf{z})$ of a dataset $\mathbf{z}$ is the value that minimises the maximum absolute deviation $MR(\mathbf{z}) = \arg\min_{\lambda}(\max_{j} |z_j - \lambda|)$. These definitions can be generalised to any L^p-norm [13] (Chapter 3).

Definition 1. *Let $\mathbf{z} = \{z_1, \ldots, z_D\}$ be a dataset and let $p \geq 1$; furthermore, let μ_p be the **p-measure of central tendency** that minimises the total p-deviation function $TD_p(\lambda)$:*

$$\mu_p = \arg\min_{\lambda} \left\{ TD_p(\lambda) = \|\mathbf{z} - \boldsymbol{\Lambda}\|_p^p = \frac{1}{D} \sum_{j=1}^{D} (z_j - \lambda)^p \right\},$$

where $\boldsymbol{\Lambda} = (\lambda, \ldots, \lambda)$, $\lambda \in \mathbb{R}$.

With this definition, the median ($\mu_1 = Med(\mathbf{z})$), arithmetic mean ($\mu_2 = \overline{\mathbf{z}}$), and mid-range ($\mu_\infty = MR(\mathbf{z})$) follow as special cases for the norms L^1, L^2, and L^∞, respectively.

Remark 1. *Convexity of the total p-deviation function $TD_p(\lambda)$:*

- *For $p = 1$, the average absolute deviation $TD_1(\lambda)$ is a convex function of λ; however, it is not strictly convex. Thus, the median may be a non-unique value.*

- *For $p > 1$, the total p-deviation function $TD_p(\lambda)$ is strictly convex; thus, if $\mu_p(\mathbf{Z})$ exists, this is unique.*

3. L^p-Norms on the Compositional Space

To define induced L^p-norms on the compositional space (CoDa L^p-norms) in a compatible way with the Aitchison geometry, one must capture the geometric structure of the $\mathcal{S}^D$ [14]. To achieve this objective, we initially define the induced L^p-norm within the quotient space $\mathcal{L}^D = \{\mathbf{z} + \lambda \mathbf{1}_D | \ \mathbf{z} \in \mathbb{R}^D, \ \mathbf{1}_D = (1, \ldots, 1)\}$. Following Brezis [15] (Chapter 11.2), an induced L^p-norm on the quotient space $\mathcal{L}^D$ can be defined by inducing the Euclidean L^p-norm in $\mathbb{R}^D$ on $\mathcal{L}^D$. The underlying idea is to assign to an equivalence class the minimum value of the L^p-norm among the elements belonging to the same equivalence class.

Definition 2. *Let $\mathbf{z} \in \mathcal{L}^D$ be a log-composition. The* **induced L^p-norm**, *denoted by $\|\mathbf{z}\|_{p,\mathcal{L}^D}$, is*

$$\|\mathbf{z}\|_{p,\mathcal{L}^D} = \min_{\lambda} \|\mathbf{z} + \lambda \mathbf{1}_D\|_p,$$

where $\mathbf{1}_D = (1, \ldots, 1)$ and $\|\cdot\|_p$ denotes the typical L^p-norm in the real space.

Using the logarithmic isomorphism [14], the L^p-norm can be extended to the compositional space.

Definition 3. *Let $\mathbf{x} \in \mathcal{S}^D$ be a composition. The* **CoDa L^p-norm**, *denoted by $\|\mathbf{x}\|_{p,\mathcal{S}^D}$, is*

$$\|\mathbf{x}\|_{p,\mathcal{S}^D} = \|\ln \mathbf{x}\|_{p,\mathcal{L}^D} = \min_{\lambda} \|\ln \mathbf{x} + \lambda \mathbf{1}_D\|_p.$$

Proposition 1. *The CoDa L^p-norm on $\mathcal{S}^D$ is $\|\mathbf{x}\|_{p,\mathcal{S}^D}$, and verifies the properties of the Aitchison geometry [16]:*

- *Scale invariance: $\|\mathbf{x}\|_{p,\mathcal{S}^D} = \|k\mathbf{x}\|_{p,\mathcal{S}^D}, k > 0$.*
- *Permutation invariance: $\|(x_1, \ldots, x_i, \ldots, x_j, \ldots, x_D)\|_{p,\mathcal{S}^D} = \|(x_1, \ldots, x_j, \ldots, x_i, \ldots, x_D)\|_{p,\mathcal{S}^D}$.*
- *Subcompositional dominance: $\|\mathbf{x}\|_{p,\mathcal{S}^D} \geq \|\text{sub}(\mathbf{x})\|_{p,\mathcal{S}^d}$, where $\text{sub}(\mathbf{x}) \in \mathcal{S}^d$ denotes any subset formed by d parts of $\mathbf{x}$.*

Proof of Proposition 1. The proof directly follows from the Definition 3. □

Following Definition 3 and the measures of central tendency described in Section 2.2, the CoDa L^p-norms L^1, L^2, and L^∞ can be developed:

- The CoDa L^1-norm on $\mathcal{S}^D$ is

$$\|\mathbf{x}\|_{1,\mathcal{S}^D} = \|\ln \mathbf{x} - Med(\ln \mathbf{x})\mathbf{1}_D\|_1 = \left\|\ln \frac{\mathbf{x}}{Med(\mathbf{x})}\right\|_1 = \sum_{j=1}^{D} \left|\ln \frac{x_j}{Med(\mathbf{x})}\right|,$$

  where $Med(\ln \mathbf{x})$ and $Med(\mathbf{x})$ are the median of the sets $\{\ln x_1, \ldots, \ln x_D\}$ and $\{x_1, \ldots, x_D\}$, respectively. As the logarithm function is strictly increasing, as per Definition 1, the set of points that serve as solutions to the variational problem TD_1 when applied to log-transformed values $\mu_1 = Med(\ln(x))$ precisely corresponds to the log-transformed set of points that are solutions to the variational problem TD_1 when applied to the raw data, that is, $\ln(Med(x))$.
  Wu et al. [17] proposed the median of a D-part composition as an alternative denominator to the geometric mean in an attempt to extend the definition of clr-scores. In general, the performance of the median as a robust estimator of the midpoint of a dataset is better when the data have high asymmetry. The CoDa L^1-norm captures the distance between two points when movement is restricted to paths that run parallel to the clr-axes ($\ln(\frac{x_i}{g(\mathbf{x})})$), as is the case in a grid or city street network (Manhattan distance, Figure 1). The CoDa L^1-norm has an equivalent expression that captures

the information about the ratio between the components of a composition; indeed, the median is the central point that divides a set into two equal parts, with half of the values falling below this central position and half above it. Therefore, half of the log-ratios $\ln(\frac{x_j}{Med(\mathbf{x})})$ are positive and the other half are negative. If we rearrange the parts of a composition in increasing order (small to large), i.e., $x_{(1)} \leq \ldots \leq x_{(D)}$, then the CoDa L^1-norm can be written in the following manner:

* $\|\mathbf{x}\|_{1,\mathcal{S}^D} = \ln\left(\frac{x_{(n+1)} \cdot \ldots \cdot x_{(2n)}}{x_{(1)} \cdot \ldots \cdot x_{(n)}}\right)$ if $D = 2n$;
* $\|\mathbf{x}\|_{1,\mathcal{S}^D} = \ln\left(\frac{x_{(n+1)} \cdot \ldots \cdot x_{(2n-1)}}{x_{(1)} \cdot \ldots \cdot x_{(n-1)}}\right)$ if $D = 2n - 1$.

Thus, the CoDa L^1-norm is a *balance* between the large parts and small parts.

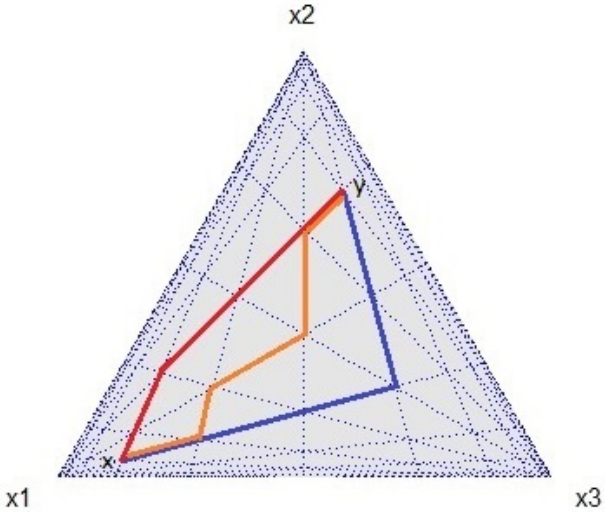


Figure 1. The Manhattan distance based on the CoDa L^1-norm in the simplex $\mathcal{S}^3$: the distance between two points $\mathbf{x} = \mathcal{C}[e^3, 1, e]$ and $\mathbf{y} = \mathcal{C}[1, e^2, e]$ in a grid-based system, where $\mathcal{C}[\cdot]$ is the closure operation, is represented by three paths (red, orange, and blue) of the same length (five units).

- The CoDa L^2-norm on $\mathcal{S}^D$ is

$$\|\mathbf{x}\|_{2,\mathcal{S}^D} = \|\ln\mathbf{x} - \overline{\ln(\mathbf{x})}\mathbf{1}_D\|_2 = \|\ln\mathbf{x} - \ln(g(\mathbf{x}))\mathbf{1}_D\|_2 = \left\|\ln\frac{\mathbf{x}}{g(\mathbf{x})}\right\|_2 = \left(\sum_{j=1}^{D}\left(\ln\frac{x_j}{g(\mathbf{x})}\right)^2\right)^{\frac{1}{2}},$$

where $g(\mathbf{x})$ is the geometric mean of the set $\{x_1, \ldots, x_D\}$. Because $\ln\frac{\mathbf{x}}{g(\mathbf{x})} \in$ clr-subspace, the CoDa L^2-norm is the restricted Euclidean L^2-norm on the clr-subspace. This norm is commonly referred to as Aitchison's norm $\|\mathbf{x}\|_{\mathcal{A}}$ [18].

- The CoDa L^∞-norm on $\mathcal{S}^D$ is

$$\|\mathbf{x}\|_{\infty,\mathcal{S}^D} = \|\ln\mathbf{x} - MR(\ln\mathbf{x})\mathbf{1}_D\|_\infty = \left\|\ln\frac{\mathbf{x}}{GR(\mathbf{x})}\right\|_\infty = \max_j\left\{\left|\ln\frac{x_j}{GR(\mathbf{x})}\right|\right\},$$

where $MR(\ln\mathbf{x})$ and $GR(\mathbf{x})$ are respectively the mid-range and geometric mid-range of the sets $\{\ln x_1, \ldots, \ln x_D\}$ and $\{x_1, \ldots, x_D\}$. Note that $MR(\ln(\mathbf{x})) = \ln(GR(\mathbf{x}))$; thus, $GR(\mathbf{x}) = \left(\max_i\{x_i\} \cdot \min_j\{x_j\}\right)^{\frac{1}{2}}$, $i, j = 1, \ldots, D$. The CoDa L^∞-norm can be interpreted as a form of log-pairwise, as the CoDa L^∞-norm represents half of the log-pairwise between the largest part against the smallest part. This log-pairwise is the greatest among all log-pairwise in the composition:

$$\|\mathbf{x}\|_{\infty,\mathcal{S}^D} = \frac{1}{2}\ln\left(\frac{\max\limits_i\{x_i\}}{\min\limits_j\{x_j\}}\right) = \frac{1}{2}\max_{i,j}\left\{\ln\frac{x_i}{x_j}\right\}.$$

4. Penalised Regression with a Compositional Covariate

The LASSO regression model is formulated as the combination of the L^2-norm cost function and the L^1-norm regularisation term [1]. For a real dataset $\mathbf{Z}$ with n observations and D predictors along with a real response vector $\mathbf{Y}$ of length n, the LASSO regression model can be formulated as follows:

$$\min\left\{\frac{1}{2}\|\mathbf{Y} - \beta_0 - \langle\boldsymbol{\beta},\ \mathbf{Z}\rangle_E\|_2^2 + \lambda\|\boldsymbol{\beta}\|_1\right\} \tag{5}$$

where β_0 is the intercept, the vector $\boldsymbol{\beta}$ is the gradient, and λ is the penalty parameter that controls the amount of regularisation. Note that $\|\cdot\|_2$ and $\|\cdot\|_1$ refer to the Euclidean L^2 and L^1 norms in real space, respectively. For $\lambda = 0$, the LASSO regression model (Equation (5)) provides the classical least squares regression model. The larger the value of λ, the greater the number of coefficients in $\boldsymbol{\beta}$ that is forced to be zero. The *optimal* value of λ can be chosen based on cross-validation techniques and related methods [19].

In the case of CoDa, additional considerations must be taken into account in order to respect the compositional nature of both the covariate $\mathbf{X}$ and the intercept $\boldsymbol{\beta}$. In variable selection, [3,20] wrote the LASSO model in terms of $\boldsymbol{\beta}^* = \ln\boldsymbol{\beta}$ and the log-transformed data instead of the clr-scores; consequently, a linear constraint on the compositional gradient coefficient $\boldsymbol{\beta}^*$ is necessary:

$$\min\left\{\frac{1}{2}\|\mathbf{Y} - \beta_0 - \langle\boldsymbol{\beta}^*,\ \ln(\mathbf{X})\rangle_E\|_2^2 + \lambda\|\boldsymbol{\beta}^*\|_1\right\}, \text{subject to } \sum_{j=1}^{D}\beta_j^* = 0. \tag{6}$$

Most of the literature addressing the topic of penalised regression with a compositional covariate has predominantly employed the Euclidean L^1-norm in the penalty term, leading to a clr-variable selection ([3–6,20–22]).

In Equation (6), the constraint $\sum_{j=1}^{D}\beta_j^* = 0$ can be incorporated in the minimising function. The constraint $\sum_{j=1}^{D}\beta_j^* = 0$ forces the $\boldsymbol{\beta}^*$ parameter to be an element in the clr-subspace. Therefore, per Equation (4), the inner product $\langle\boldsymbol{\beta}^*,\ \ln(\mathbf{X})\rangle_E$ is equivalent to $\langle\mathrm{clr}(\boldsymbol{\beta}),\ \mathrm{clr}(\mathbf{X})\rangle_E = \langle\boldsymbol{\beta},\ \mathbf{X}\rangle_{\mathcal{A}}$. Thus, the constrained LASSO (Equation (6)) is equivalent to the following definition.

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

$$\boldsymbol{\beta} \in \underset{\beta_0,\,\boldsymbol{\beta}}{\operatorname{argmin}}\left\{\frac{1}{2}\|\mathbf{Y} - \beta_0 - \langle\boldsymbol{\beta},\ \mathbf{X}\rangle_{\mathcal{A}}\|_2^2 + \lambda\|\boldsymbol{\beta}\|_{1-clr}\right\}, \tag{7}$$

where $\|\boldsymbol{\beta}\|_{1-clr} = \sum_{j=1}^{D}|\mathrm{clr}(\boldsymbol{\beta})_j| = \sum_{j=1}^{D}\left|\ln\frac{\beta_j}{g(\boldsymbol{\beta})}\right|$.

The key innovation here is that the linear constraint becomes embedded in the penalty term through the L^1-clr norm. This change in approach is not merely an algebraic or formal change; rather it implies a deeper understanding of the variable selection process in CoDa. The penalty term imposes a constraint on the sum of the absolute values of clr-scores within the gradient vector $\boldsymbol{\beta}$. This constraint compels the model to shrink or eliminate certain clr-scores, effectively driving them to zero. Consequently, this results in a balance selection. Without loss of generality, let us assume that the balance $\ln\frac{\beta_1}{g(\boldsymbol{\beta})}$ is zero. This implies that the corresponding balance $\ln\frac{x_1}{g(\mathbf{x})}$ has no influence on the response variable y. Therefore,

the maximum variation in y is concentrated in the subspace orthogonal to the balance $\ln \frac{x_1}{g(\mathbf{x})}$, i.e., the subspace of balances among the subcomposition $(x_2, \ldots, x_D)$. This selective regularization process facilitates variable selection, as x_1 does not influence the response variable y.

In order to establish a unified framework, the generalised LASSO problem [23] can be adapted to penalised linear models with a compositional covariate:

$$\min \left\{ \frac{1}{2} \|\mathbf{Y} - \beta_0 - \langle \boldsymbol{\beta}^*, \operatorname{clr}(\mathbf{X}) \rangle_E \|_2^2 + \lambda \|\mathbf{D} \cdot \boldsymbol{\beta}^*\|_1 \right\}, \tag{8}$$

where $\|\cdot\|_2$ and $\|\cdot\|_1$ respectively refer to the Euclidean L^2 and L^1 norms in real space. The generalised LASSO problem allows for a broader range of applications by considering a matrix $\mathbf{D}$ associated with the penalty term. The matrix $\mathbf{D}$ is related to the L^1-norm considered in the penalty term. The choice of one norm over another determines the type of regularization. Different models can be formulated within the framework of the generalised LASSO problem and addressed through convex optimization algorithms. Solving each of these different penalised regression models yields distinct coefficients, each characterised by unique properties. Indeed, Definition 4 can be expressed as a generalised LASSO problem in the following manner:

$$\boldsymbol{\beta}^* \in \underset{\beta_0,\, \boldsymbol{\beta}^*}{\operatorname{argmin}} \left\{ \frac{1}{2} \sum_{i=1}^{n} (y_i - \beta_0 - \langle \boldsymbol{\beta}^*, \operatorname{clr}(\mathbf{X})_i \rangle_E)^2 + \lambda \|\mathbf{C}_D \boldsymbol{\beta}^*\|_1 \right\}, \tag{9}$$

where $\boldsymbol{\beta}^* = \ln \boldsymbol{\beta}$ and $\mathbf{D} = \mathbf{C}_D$ is the centering matrix on the clr-subspace, with $\mathbf{C}_D \boldsymbol{\beta}^* = \boldsymbol{\beta}^* - \overline{\boldsymbol{\beta}^*} \mathbf{1}_D$.

On the other hand, following [7], it is possible to consider the matrix $\mathbf{D}$ equal to $\mathbf{F}$, that is, the matrix associated with the linear transformation $F(\beta_1^*, \ldots, \beta_D^*) = \frac{1}{D-1}(\beta_1^* - \beta_2^*, \beta_1^* - \beta_3^*, \ldots, \beta_1^* - \beta_D^*, \beta_2^* - \beta_3^*, \ldots, \beta_2^* - \beta_D^*, \ldots, \beta_{D-1}^* - \beta_D^*)$. Note that $\beta_i^* - \beta_j^* = \ln\left(\frac{\beta_i}{\beta_j}\right)$, which is a log-pairwise. In this case, the penalty term in a generalised LASSO problem can be written as $\|\mathbf{F} \cdot \boldsymbol{\beta}^*\|_1$, meaning that the generalised LASSO problem results in the following:

$$\boldsymbol{\beta}^* \in \underset{\beta_0,\, \boldsymbol{\beta}^*}{\operatorname{argmin}} \left\{ \frac{1}{2} \sum_{i=1}^{n} (y_i - \beta_0 - \langle \boldsymbol{\beta}^*, \operatorname{clr}(\mathbf{X})_i \rangle_E)^2 + \lambda \|\mathbf{F} \boldsymbol{\beta}^*\|_1 \right\}. \tag{10}$$

The model can be defined as follows.

Definition 5. *Given y_i, $i = 1, \ldots, n$, the sample of the response variable $\mathbf{X}$, and the $n \times D$ matrix whose rows, $\mathbf{X}_i = (x_{i1}, \ldots, x_{iD})$ for $i = 1, \ldots, n$ contain the compositional sample, the L^1-plr LASSO estimator is defined as*

$$\boldsymbol{\beta} \in \underset{\beta_0,\, \boldsymbol{\beta}}{\operatorname{argmin}} \left\{ \frac{1}{2} \|\mathbf{Y} - \beta_0 - \langle \boldsymbol{\beta}, \mathbf{X} \rangle_{\mathcal{A}} \|_2^2 + \lambda \|\boldsymbol{\beta}\|_{1-plr} \right\}, \tag{11}$$

where $\|\boldsymbol{\beta}\|_{1-plr} = \frac{1}{D-1} \sum_{i<j} \left| \ln\left(\frac{\beta_i}{\beta_j}\right) \right|$.

Importantly, because $\ln \frac{\beta_i}{\beta_j} = \operatorname{clr}(\boldsymbol{\beta})_i - \operatorname{clr}(\boldsymbol{\beta})_j$, the penalty term shrinks the absolute value of the differences of the clr-scores within the gradient vector, which forces some pairwise differences of clr-scores to be zero, i.e., it forces equality on some clr-scores. Therefore, each set of equal clr-scores defines a subcomposition with non-influential balances within its parts. This selective regularization process facilitates balanced selection [7].

Finally, using the CoDa L^1-norm introduced in Section 3, it is possible to define another generalised LASSO problem.

Definition 6. *Given y_i, $i = 1, \ldots, n$, the sample of the response variable* $\mathbf{X}$*, and the $n \times D$ matrix whose rows $\mathbf{X}_i = (x_{i1}, \ldots, x_{iD})$ for $i = 1, \ldots, n$ contain the compositional sample, the CoDa L^1-norm LASSO estimator is defined as*

$$\boldsymbol{\beta} \in \underset{\beta_0,\, \boldsymbol{\beta}}{\operatorname{argmin}} \left\{ \frac{1}{2} \|\mathbf{Y} - \beta_0 - \langle \boldsymbol{\beta},\, \mathbf{X} \rangle_{\mathcal{A}}\|_2^2 + \lambda \|\boldsymbol{\beta}\|_{1,\mathcal{S}^D} \right\}, \tag{12}$$

where $\|\boldsymbol{\beta}\|_{1,\mathcal{S}^D} = \sum_{j=1}^{D} \left| \ln \frac{\beta_j}{Med(\boldsymbol{\beta})} \right|$.

In this case, the penalty term compels certain parts β_j to be equal to the median of the parts, ensuring equality among them in particular. Consequently, the effect produced is also a balance selection, as in the previous case; however, unlike the L^1-plr LASSO estimator, with the CoDa L^1-norm estimator there is only one set of equal clr-scores, and all non-influential balances belong to a single subcomposition.

As there is no algebraic formula to express the median, $Med(\boldsymbol{\beta})$, it is necessary to include a new variable $m \in \mathbb{R}$ in the penalty term when formulating the minimization problem (Equation (12)) as a generalised LASSO problem:

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

where $\boldsymbol{\beta}^* = \ln \boldsymbol{\beta}$, $\mathbf{D} = \mathbf{G}$, and the matrix associated with the transformation $G(\boldsymbol{\beta}^*, m) = (\boldsymbol{\beta}_1^* - m, \boldsymbol{\beta}_2^* - m, \ldots, \boldsymbol{\beta}_D^* - m)$.

5. Study Case

We used the microbial dataset analyzed in [24,25] to compare the different L^1-norms in a CoDa LASSO regression problem. The dataset, collected and explained in [24], comprises the compositions of $D = 60$ taxa spanning various taxonomic levels (e.g., g for genus, f for family, o for order, and k for kingdom) within a set of n =151 individuals. The dependent variable y is an inflammatory parameter, specifically, the levels of soluble CD14 (sCD14 variable) measured for each individual. These data are available in the R package [26]. An individual having a zero value recorded for some parts indicates that certain taxa were not detected. A zero value prevents the application of the log-ratio methodology. Following a more analogous procedure than in [26], the genus *Thalassospira*, *unclassified* genus of the class *Alphaproteobacteria*, and *unclassified* genus of the family *Porphyromonadaceae*, all with more than 80% of zeros, were removed. The rest of the zeros recorded in the remaining 57 taxa were replaced by a small value using an imputation method [27,28]. Because the zeros are of count type, it is appropriate to apply methods based on Dirichlet-multinomial duality [29].

To solve the convex optimizations problems in Equations (9), (10), and (13), we first select the optimal λ parameter for the penalised model by performing a ten-fold cross-validation. Each iteration involves dividing the data into ten 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). With the parameter λ selected, we proceeded to solve the optimization problem in order to find the parameters β_0 and $\boldsymbol{\beta}^*$. The CVXR package in R version 4.3.2 [30] offers an interface for defining and solving convex optimization problems. CVXR utilises a domain-specific language, making it user-friendly and allowing users to express optimization problems. The package supports various solvers, enabling users to choose the one that best suits their needs. In our case, we opted for the Operator Splitting Quadratic Program (OSQP). The OSQP is a solver for quadratic programming problems and employs an operator-splitting method [31]. This solver is highly efficient even in cases where the matrices are not full-rank, such as our situation, because the clr-scores are used. Referring to the procedure detailed above, we have outlined an Algorithm 1 for a generalised LASSO method below.

The algorithm is applied in the three cases discussed in the previous section, namely, when considering the three different L^1 norms in the penalty term, i.e., the L^1-clr (Definition 4), L^1-plr (Definition 5), and CoDa L^1-norms (Definition 6).

For the L^1-clr estimator, the LASSO regression algorithm is applied iteratively within the cross-validation framework. Figure 2 illustrates the cross-validated MSE across different λ values. The optimal λ is determined by selecting the point on the curve where the mean squared error is minimised: λ_{min} = 35,769.42.

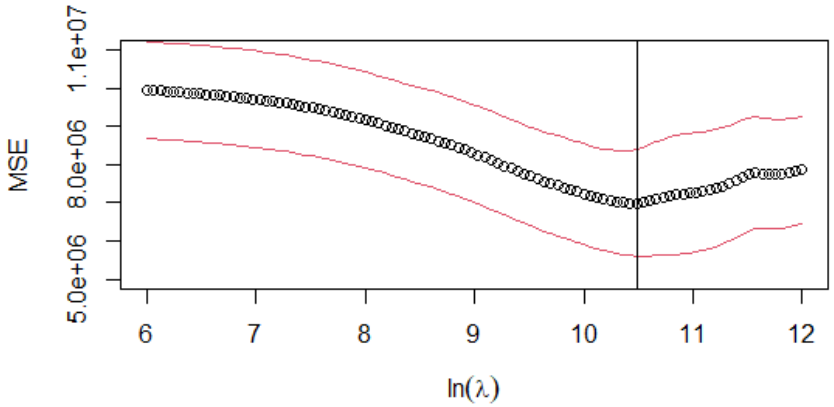


Figure 2. L^1-clr: cross-validation MSE curve for different log-transformed values of the penalty parameter ($\ln(\lambda)$). The circle ($\circ$) is the arithmetical mean of the ten-fold CV. The red lines (above and below the mean) indicate the mean $\pm$ stdev value, where stdev is the standard deviation of the ten-fold CV. The vertical line represents the log-transformed values of lambda.min = 35,769.42.

Algorithm 1 Generalised LASSO for CoDa

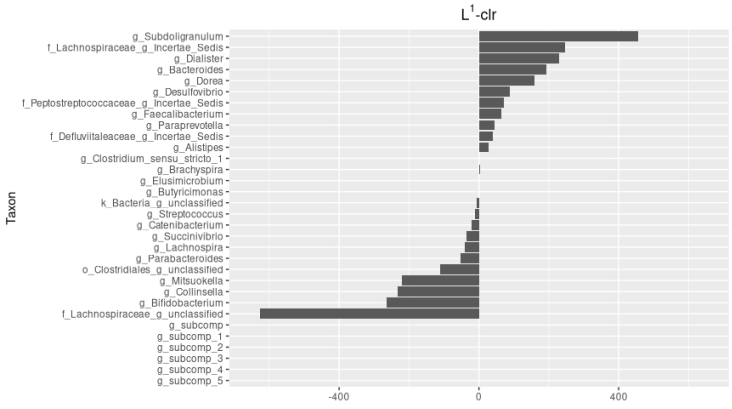
1. Fit the generalised LASSO model with tuning parameter λ (Equations (9), (10) or (13)).
2. Calculate the clr-representative: $\boldsymbol{\beta}^* - \overline{\boldsymbol{\beta}^*}$
3. Express the generalised LASSO model in terms of clr-scores.

For λ_{min} = 35,769.42, the generalised LASSO (Equation (9)) identifies which ones among all the β_i^* are set to $\overline{\boldsymbol{\beta}^*}$, particularly ensuring equality among them. Importantly, when computing the representative $\boldsymbol{\beta}^* - \overline{\boldsymbol{\beta}^*}\mathbf{1}_D \in$ clr-subspace, we find that some clr-scores are equal to zero. Therefore, the regularization process effectively splits the composition into two subcompositions. The first subcomposition represents the 33 non-influential parts, where coefficients $\text{clr}(\boldsymbol{\beta})_k$, $k = 1, \ldots, 33$ are driven to zero, contributing to model simplicity. The second subcomposition identifies the 24 parts that actively contribute to the influential balances on the response variable y (see Table A1). The intercept β_0 is equal to 6563.19. Figure 3a shows the non-zero clr-scores for parameter $\boldsymbol{\beta}$. We highlight that the most influential pairwise is formed by the genus *Subdoligranulum* and the unclassified genus of the family *Lachnospiraceae*.

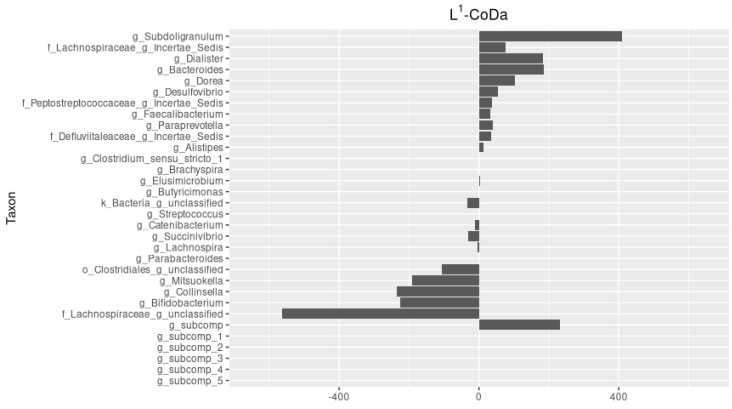
For the CoDa L^1-norm estimator (Equation (13)), the LASSO regression algorithm is applied with the same cross-validation partition used in the L^1-clr estimator. Figure 4 illustrates the model's performance across different regularization parameters λ. The optimal value is λ_{min} = 45,582.21

For λ_{min} = 45,582.21, the generalised LASSO (Equation (13)) identifies which ones among all the β_i^* are set to the median of $\boldsymbol{\beta}^*$, particularly ensuring equality among them. This equality among some $\boldsymbol{\beta}_i^*$ indicates that the balances involving their respective parts x_i have a non-influential role in the response variable y. However, in contrast to the L^1-clr scenario when computing the representative $\boldsymbol{\beta}^* - \overline{\boldsymbol{\beta}^*}$, in general, all clr-scores are non-zero. Consequently, variable selection cannot be performed in this case. The regularization process effectively splits the composition into two subcompositions. The first subcomposition

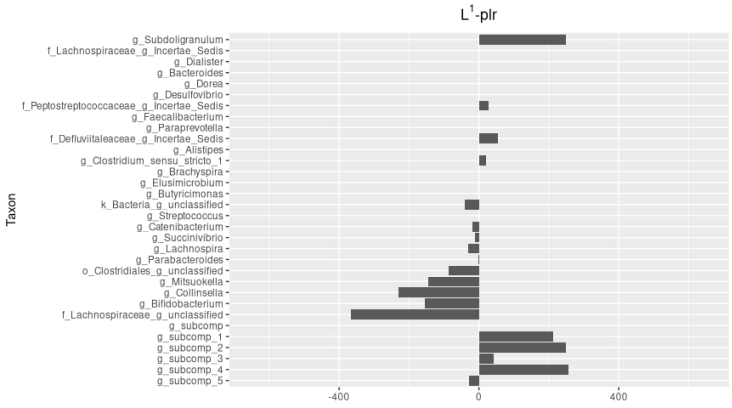
represents the 36 *internally independent* parts, that is, a subcomposition in which the balances between the respective parts do not influence y [32]. The coefficients $\mathrm{clr}(\boldsymbol{\beta})_k$, $k = 1, \ldots, 36$ are driven to the median of $\boldsymbol{\beta}^*$ (6.44), contributing to model simplicity. The second subcomposition identifies the 21 parts that actively contribute to the influential balances on the response variable y (see Table A1).



(**a**)



(**b**)



(**c**)

Figure 3. Comparison of the $\mathrm{clr}(\boldsymbol{\beta})$ parameter, with the taxon order maintained on the vertical axis to facilitate comparison: (**a**) clr-scores for the L^1-clr LASSO estimator, (**b**) clr-scores for the L^1-CoDa LASSO estimator, and (**c**) clr-scores for the L^1-plr LASSO estimator.

To highlight the model's simplicity, it is crucial to accurately summarise the information contained in the first subcomposition. Without loss of generality, let $(x_1, \ldots, x_k)$ be an internally independent subcomposition. The linear model in clr-scores is

$$y = \beta_0 + \sum_{j=1}^{k} \operatorname{clr}(\beta)_j \ln x_i + \sum_{j=k+1}^{D} \operatorname{clr}(\beta)_j \ln x_j,\ \operatorname{clr}(\beta)_1 = \ldots = \operatorname{clr}(\beta)_k.$$

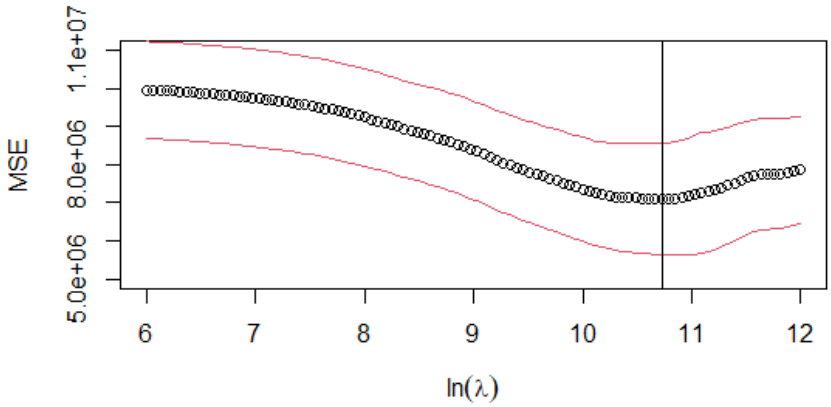


Figure 4. CoDa L^1-norm: cross-validation MSE curve for different log-transformed values of the penalty parameter ($\ln(\lambda)$). The circle ($\circ$) is the arithmetical mean of the ten-fold CV. The red lines (above and below the mean) represent the value mean $\pm$ stdev, where stdev is the standard deviation of the ten-fold CV. The vertical line represents the log-transformed values of lambda.min = 45,582.21.

As explained by [16] in Chapter 4, the best approach to represent a subcomposition is through its geometric mean, denoted as $g_{sub} = (x_1 \cdot \ldots \cdot x_k)^{\frac{1}{k}}$. Therefore, the linear model is

$$y = \beta_0 + k\operatorname{clr}(\beta_{sub}) \ln g_{sub} + \sum_{j=k+1}^{D} \operatorname{clr}(\beta)_j \ln x_j,$$

where $\operatorname{clr}(\beta_{sub}) = \operatorname{clr}(\beta)_1 = \ldots = \operatorname{clr}(\beta)_k$. This model has $D - k$ degrees of freedom, as opposed to the $D - 1$ degrees of freedom of the general linear model. The intercept value is $\beta_0 = 7023.879$, and Figure 3b shows the clr-scores of $\boldsymbol{\beta}$.

L^1-clr regularization creates a subcomposition that is both internally and externally independent [32], that is, both the balances within the parts of the subcomposition and the full balance between the parts of the subcomposition and the rest of the parts are all non-influential. In contrast, CoDa L^1-norm regularization relaxes the conditions and establishes only one subcomposition that is internally independent. In this context, the CoDa L^1-norm is somewhat more permissive. When comparing the results, we observe that both are quite similar; what stands out is the significance of the new variable g_{sub} in the CoDa L^1-norm penalised linear model. L^1-clr regularization eliminates the balance $\ln \frac{g_{sub}}{g(\mathbf{x})}$ without prior analysis. This observation prompts us to consider that the direct application of L^1-clr regularization might be premature. Furthermore, when dealing with a penalised model, it is always possible to subsequently test the nullity of any parameter [33].

L^1-clr and CoDa L^1-norm regularization share the fact that both shrink the difference between β_i^* coefficients and a central measure, respectively, the mean and the median; consequently, each regularization technique generates a unique subcomposition with certain properties related to its influence on the dependent variable y. Because the goal of a CoDa analysis is to describe the subcompositional structure of the data, the use of the L^1-clr and CoDa L^1-norms in the penalty term leads to a result that has to be considered as limited. To overcome this limitation, the L^1-plr norm enables the construction of more than one internally independent subcomposition, which can better capture the subcompositional structure of the data regarding the variable y [7].

With the same data partition as executed in previous cases, we performed cross-validation to find the optimal lambda value for the L^1-plr estimator. The optimal parameter is λ = 69,669.31 (Figure 5).

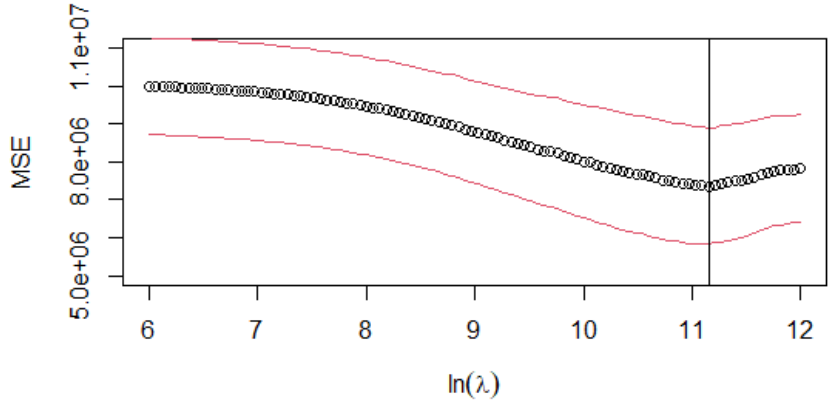


Figure 5. L^1-plr: cross-validation MSE curve for different log-transformed values of the penalty parameter ($\ln(\lambda)$). The circle ($\circ$) is the arithmetical mean of the ten-fold CV. The red lines (above and below the mean) represent the value mean $\pm$ stdev, where stdev is the standard deviation of the ten-fold CV. The vertical line represents the log-transformed values of lambda.min = 69,669.31.

For λ = 69,669.31, the generalised LASSO (Equation (13)) splits the composition into six distinct subcompositions: five internally independent subcompositions on response variable y and one subcomposition comprising 14 parts actively contributing to the influential balances on the response variable y (see Table A1 to compare L^1-plr estimator with L^1-clr and L^1-CoDa estimators, and Table A2 to explore its subcompositional structure). Each of the five internally independent subcompositions related to y contributes to reducing the dimension of the linear model. This reduction is achieved by substituting each subcomposition with its geometric mean ($g - subcomp_k$, $k = 1, \ldots, 5$), following the approach outlined in the CoDa L^1-norm estimator. The intercept value is $\beta_0 = 7023.879$ and Figure 3c shows the clr-scores of $\boldsymbol{\beta}$.

The L^1-plr estimator is the simplest and provides us with the most information about the subcompositional structure of the composition as regards the variable y.

6. Discussion

This paper has rigorously defined CoDa L^p-norms, providing a foundation for their application. The specific cases of the CoDa L^1, L^2, and L^∞ norms have been studied, interpreting these metrics in terms of log-ratios to enhance the reader's understanding. Additionally, a unified treatment of three distinct L^1-norms tailored for compositional data has been presented in the context of a generalised LASSO problem. Through a detailed examination of the regularization effects of each norm, we have uncovered valuable insights. The L^1-clr norm is well suited for variable selection, creating a unique subcomposition that is both internally and externally independent. The CoDa L^1-norm, on the other hand, emphasises internal independence. Lastly, the L^1-plr norm showcases a balance selection effect. Consequently, the L^1-plr norm enables more detailed study of the subcompositional structure of the compositional covariate $\mathbf{x}$ in relation to the explained variable y.

In this article, we have expanded the methodological toolkit for performing penalised regression with compositional covariates. For low dimensions, our recommendation is to run penalised regression with the L^1-plr norm. However, we cannot ignore that variable selection becomes imperative for higher dimensions. Therefore, we suggest conducting an initial examination using the CoDa L^1-norm or L^1-plr norm to gain insights into the subcompositional structure. Following this analysis, it is possible to proceed with penalised regression employing the L^1-clr norm.

As part of our future work, we aim to investigate penalised regression models that effectively integrate both the L^1-plr and L^1-clr norms into the penalty term. This research is expected to offer deeper insight into the underlying structure of compositional data, allowing for a more thorough understanding. Moreover, our aim is to improve the flexibility of modelling, especially in datasets with high dimensionality. This holistic approach will contribute to advancing the applicability and effectiveness of penalised regression techniques in the context of compositional data analysis.

Author Contributions: Conceptualization, J.A.M.-F., J.S.-R. and G.M.-F.; Formal analysis, J.A.M.-F., J.S.-R. and G.M.-F.; Methodology, J.A.M.-F., J.S.-R. and G.M.-F.; Software, J.S.-R.; Supervision, J.A.M.-F. and G.M.-F. All authors have read and agreed to the published version of the manuscript.

Funding: This research was funded by Agency for Administration of University and Research grant number 2021SGR01197, and Ministerio de Ciencia e Innovación grant number PID2021-123833OB-I00, and Ministerio de Ciencia e Innovación grant number PRE2019-090976.

Data Availability Statement: Data are contained within the article.

Conflicts of Interest: The authors declare no conflicts of interest.

Abbreviations

The following abbreviations are used in this manuscript:

LASSO	Least Absolute Shrinkage and Selection Operator
CoDa	Compositional Data
clr	Centered Log-Ratio
L^1-plr	Pairwise Log-Ration norm
L^1-clr	Centered Log-Ration norm
TD	Total p-Deviation Function
MSE	Mean Squarred Error
OSQP	Operator-Splitting Quadratic Program

Appendix A

Table A1. clr-scores for the three LASSO estimators grouped into subcompositions.

Taxon	L1-clr	L1-CoDa	L1-plr
Intercept	6563.19	7023.88	7244.88
g_Subdoligranulum	455.51	409.93	249.27
f_Lachnospiraceae_g_Incertae_Sedis	245.69	77.35	
g_Dialister	229.23	182.53	
g_Bacteroides	193.46	186.23	
g_Dorea	159.97	104.06	
g_Desulfovibrio	88.32	53.57	
f_Peptostreptococcaceae_g_Incertae_Sedis	70.37	37.38	26.54
g_Faecalibacterium	63.74	33.07	
g_Paraprevotella	45.15	38.71	
f_Defluviitaleaceae_g_Incertae_Sedis	39.35	34.48	54.86
g_Alistipes	27.18	14.05	
g_Clostridium_sensu_stricto_1			20.52
g_Brachyspira	4.33		
g_Elusimicrobium		2.25	
g_Butyricimonas	0.16		
k_Bacteria_g_unclassified	−7.48	−33.79	−41.56
g_Streptococcus	−11.71		
g_Catenibacterium	−21.83	−12.10	−19.66
g_Succinivibrio	−34.64	−31.51	−10.83
g_Lachnospira	−40.80	−4.21	−29.65

Table A1. *Cont.*

Taxon	L1-clr	L1-CoDa	L1-plr
g_Parabacteroides	−52.41		−0.28
o_Clostridiales_g_unclassified	−111.82	−107.08	−85.66
g_Mitsuokella	−219.27	−192.13	−144.62
g_Collinsella	−233.08	−235.68	−228.94
g_Bifidobacterium	−263.23	−225.95	−154.34
f_Lachnospiraceae_g_unclassified	−626.20	−563.00	−365.84
g_subcomp		231.83	
g_subcomp_1			212.52
g_subcomp_2			248.34
g_subcomp_3			41.25
g_subcomp_4			255.78
g_subcomp_5			−27.66

Table A2. Details of the subcompositional structure for the L^1-plr LASSO estimator.

Taxon	clr(β)
g_Subdoligranulum	249.27
g_subcomp_1: g_Bacteroides, g_Dialister	212.52
f_Defluviitaleaceae_g_Incertae_Sedis	54.86
g_subcomp_2: f_Lachnospiraceae_g_Incertae_Sedis, g_Dorea, g_Faecalibacterium, g_Alistipes, g_Desulfovibrio, g_Paraprevotella	248.34
f_Peptostreptococcaceae_g_Incertae_Sedis	26.54
g_Clostridium_sensu_stricto_1	20.52
g_subcomp_3: g_Escherichia-Shigella, f_Ruminococcaceae_g_unclassified, g_Butyricimonas	41.25
g_subcomp_4: g_Brachyspira, g_Barnesiella, g_Blautia, f_Rikenellaceae_g_unclassified, g_Odoribacter, f_Erysipelotrichaceae_g_unclassified, g_Streptococcus, g_Anaerostipes, g_Phascolarctobacterium, g_Acidaminococcus, g_Anaerovibrio, g_Roseburia, g_Alloprevotella, f_Erysipelotrichaceae_g_Incertae_Sedis, g_Megasphaera, g_Coprococcus, g_Intestinimonas, g_Solobacterium, g_Oribacterium, g_Anaeroplasma, g_Victivallis, f_Ruminococcaceae_g_Incertae_Sedis, o_NB1-n_g_unclassified, g_Sutterella, o_Bacteroidales_g_unclassified, g_Prevotella, g_RC9_gut_group, f_Christensenellaceae_g_unclassified, g_Anaerotruncus	255.78
g_Parabacteroides	−0.28
g_subcomp_5: g_Ruminococcus, g_Elusimicrobium, f_vadinBB60_g_unclassified	−27.66
g_Succinivibrio	−10.83
g_Catenibacterium	−19.66
g_Lachnospira	−29.65
k_Bacteria_g_unclassified	−41.56
o_Clostridiales_g_unclassified	−85.66
g_Mitsuokella	−144.62
g_Bifidobacterium	−154.34
g_Collinsella	−228.94
f_Lachnospiraceae_g_unclassified	−365.84

References

1. Tibshirani, R. Regression shrinkage and selection via the lasso. *J. R. Stat. Soc. Ser. B* **1996**, *58*, 267–288. [CrossRef]
2. Aitchison, J. *The Statistical Analysis of Compositional Data*; Chapman & Hall: London, UK, 1986.
3. Lin, W.; Shi, R.; Feng, R.; Li, H. Variable selection in regression with compositional covariates. *Biometrika* **2014**, *101*, 785–797. [CrossRef]
4. Shi, P.; Zhang, A.; Li, H. Regression analysis for microbiome compositional data. *Ann. Appl. Stat.* **2016**, *10*, 1019–1040. [CrossRef]
5. Lu, J.; Shi, P.; Li, H. Generalized linear models with linear constraints for microbiome compositional data. *Biometrics* **2019**, *75*, 235–244. [CrossRef] [PubMed]
6. Susin, A.; Wang, Y.; Lê Cao, K.A.; Calle, M.L. Variable selection in microbiome compositional data analysis. *NAR Genom. Bioinform.* **2020**, *2*, lqaa029. [CrossRef] [PubMed]
7. Saperas-Riera, J.; Martín-Fernández, J.; Mateu-Figueras, G. Lasso regression method for a compositional covariate regularised by the norm L^1 pairwise logratio. *J. Geochem. Explor.* **2023**, *255*, 107327. [CrossRef]
8. Egozcue, J.J.; Pawlowsky-Glahn, V. Groups of parts and their balances in compositional data analysis. *Math. Geol.* **2005**, *37*, 795–828. [CrossRef]
9. Pawlowsky-Glahn, V.; Egozcue, J.J. Geometric approach to statistical analysis on the simplex. *Stoch. Environ. Res. Risk Assess.* **2001**, *15*, 384–398. [CrossRef]
10. Billheimer, D.; Guttorp, P.; Fagan, W.F. Statistical Interpretation of Species Composition. *J. Am. Stat. Assoc.* **2001**, *96*, 1205–1214. [CrossRef]
11. Aitchison, J.; Bacon-Shone, J. Log contrast models for experiments with mixtures. *Biometrika* **1984**, *71*, 323–330. [CrossRef]
12. Van der Boogaart, K.G.; Tolosana, R. *Analyzing Compositional Data with R*; Use R!; Springer: Berlin/Heidelberg, Germany, 2013.
13. Dave, A. Measurement of Central Tendency. In *Applied Statistics for Economics*; Horizon Press: Toronto ON, Canada, 2014; Chapter 3.
14. Barceló-Vidal, C.; Martín-Fernández, J.A. The Mathematics of Compositional Analysis. *Austrian J. Stat.* **2016**, *45*, 57–71. [CrossRef]
15. Brezis, H. *Functional Analysis, Sobolev Spaces and Partial Differential Equations*; Universitext; Springer: New York, NY, USA, 2011.
16. Pawlowsky-Glahn, V.; Egozcue, J.J.; Tolosana-Delgado, R. *Modeling and Analysis of Compositional Data*; John Wiley & Sons: Chichester, UK, 2015.
17. Wu, J.R.; Macklaim, J.M.; Genge, B.L.; Gloor, G.B. Finding the Centre: Compositional Asymmetry in High-Throughput Sequencing Datasets. In *Advances in Compositional Data Analisys*; Springer: Berlin/Heidelberg, Germany, 2021; Chapter 17, pp. 329–342. [CrossRef]
18. Martín-Fernández, J. Measures of Difference and Non-Parametric Classification of Compositional Data. Ph.D. Thesis, Department of Applied Mathematics, Universitat Politècnica de Catalunya, Barcelona, Spain, 2001.
19. James, G.; Witten, D.; Hastie, T.; Tibshirani, R. *Introduction to Statistical Learning*, 2nd ed.; Springer: New York, NY USA, 2021.
20. Bates, S.; Tibshirani, R. Log-ratio lasso: Scalable, sparse estimation for log-ratio models. *Biometrics* **2019**, *75*, 613–624. [CrossRef]
21. Monti, G.; Filzmoser, P. Sparse least trimmed squares regression with compositional covariates for high-dimensional data. *Bioinformatics* **2021**, *37*, 3805–3814. [CrossRef] [PubMed]
22. Monti, G.; Filzmoser, P. Robust logistic zero-sum regression for microbiome compositional data. *Adv. Data Anal. Classif.* **2022**, *16*, 301–324. [CrossRef]
23. Tibshirani, R.; Taylor, J. The solution path of the generalized lasso. *Ann. Statist.* **2011**, *39*, 1335–1371. [CrossRef]
24. Noguera-Julian, M.; Rocafort, M.; Guillén, Y.; Rivera, J.; Casadellà, M.; Nowak, P.; Hildebrand, F.; Zeller, G.; Parera, M.; Bellido, R.; et al. Gut Microbiota Linked to Sexual Preference and HIV Infection. *eBioMedicine* **2016**, *5*, 135–146. [CrossRef] [PubMed]
25. Rivera-Pinto, J.; Egozcue, J.J.; Pawlowsky-Glahn, V.; Paredes, R.; Noguera-Julian, M.; Calle, M.L. Balances: A new perspective for microbiome analysis. *mSystems* **2018**, *3*, e00053-18. [CrossRef] [PubMed]
26. Calle, M.; Susin, T.; Pujolassos, M. coda4microbiome: Compositional Data Analysis for Microbiome Studies; R Package Version 0.2.1. *BMC Bioinf.* **2023**, *24*, 82.
27. Palarea-Albaladejo, J.; Martín-Fernández, J.A. zCompositions—R package for multivariate imputation of left-censored data under a compositional approach. *Chemom. Intell. Lab. Syst.* **2015**, *143*, 85–96. [CrossRef]
28. Palarea-Albaladejo, J.; Martin-Fernandez, J. zCompositions: Treatment of Zeros, Left-Censored and Missing Values in Compositional Data Sets; R Package Version 1.5. 2023. Available online: https://cran.r-project.org/web/packages/zCompositions/zCompositions.pdf (accessed on 13 March 2024).
29. Martín-Fernández, J.; Hron, K.; Templ, M.; Filzmoser, P.; Palarea-Albaladejo, J. Bayesian-multiplicative treatment of count zeros in compositional data sets. *Stat. Model.* **2015**, *15*, 134–158. [CrossRef]
30. Fu, A.; Narasimhan, B.; Boyd, S. CVXR: An R Package for Disciplined Convex Optimization. *J. Stat. Softw.* **2020**, *94*, 1–34. [CrossRef]
31. Stellato, B.; Banjac, G.; Goulart, P.; Bemporad, A.; Boyd, S. OSQP: An Operator Splitting Solver for Quadratic Programs. *Math. Program. Comput.* **2020**, *12*, 637–672. [CrossRef]

32. Boogaart, K.; Filzmoser, P.; Hron, K.; Templ, M.; Tolosana-Delgado, R. Classical and robust regression analysis with compositional data. *Math. Geosci.* **2021**, *53*, 823–858. [CrossRef]
33. Hyun, S.; G'Sell, M.; Tibshirani, R.J. Exact post-selection inference for the generalized lasso path. *Electron. J. Stat.* **2018**, *12*, 1053–1097. [CrossRef]