SORT 44 (1) January-June 2020, 201-220 DOI: 10.2436/20.8080.02.100

# On interpretations of tests and effect sizes in regression models with a compositional predictor

Germà Coenders[1] and Vera Pawlowsky-Glahn[2]

**Abstract**

Compositional data analysis is concerned with the relative importance of positive variables, expressed through their log-ratios. The literature has proposed a range of manners to compute log-ratios, some of whose interrelationships have never been reported when used as explanatory variables in regression models. This article shows their similarities and differences in interpretation based on the notion that one log-ratio has to be interpreted keeping all others constant. The article shows that centred, additive, pivot, balance and pairwise log-ratios lead to simple reparametrizations of the same model which can be combined to provide useful tests and comparable effect size estimates.

*MSC:* 62J05, 62H15.

*Keywords:* Compositional regression models, CoDa, composition as explanatory, centred log-ratios, pivot coordinates, pairwise log-ratios, additive log-ratios, effect size.

## 1 Introduction

*Compositional Data* (CoDa) can be defined as positive vectors containing information about the relative importance of parts of a whole, not necessarily with a constant sum.

The CoDa tradition started with Aitchison's seminal work in Aitchison (1982), and Aitchison (1986) e.g. on chemical and geological compositions. There, only the proportion of each part or component is of interest, since absolute amounts are in general either not available or irrelevant, as they only inform about the size of the chemical or soil sample. In the last three decades, CoDa have provided a standardized toolbox for statistical analyses where the research questions concern the relative importance of magnitudes, in both hard sciences and social sciences (Coenders and Ferrer-Rosell, 2020). The term *compositional analysis* (Barceló-Vidal and Martín-Fernández, 2016) has even been coined to stress the fact that what is ultimately compositional is not the data, which

[1] Dept. of Economics, University of Girona, Faculty of Economics and Business, Campus Montilivi, 17003 Girona, Spain. E-mail: germa.coenders@udg.edu

[2] Dept. of Computer Sciences, Applied Mathematics and Statistics, University of Girona, Polytechnic IV, Campus Montilivi, 17003 Girona, Spain. E-mail: vera.pawlowsky@udg.edu

Received: July 2019
Accepted: May 2020

may even not be parts of any whole, but the research objectives, research questions or hypotheses focusing on relative importance rather than absolute values. Along similar lines, CoDa have also been defined as "arrays of strictly positive numbers for which ratios between them are considered to be relevant" without any further requirement (Egozcue and Pawlowsky-Glahn, 2019). Examples of applications to data which do not represent parts of any whole can be found in Ortells et al. (2016) and Linares-Mustarós, Coenders and Vives-Mestres, 2018. Accessible handbooks have contributed to extending the use of CoDa (Buccianti, Mateu-Figueras and Pawlowsky-Glahn, 2006; Boogaart and Tolosana-Delgado, 2013; Filzmoser, Hron and Templ, 2018; Greenacre, 2018; Pawlowsky-Glahn and Buccianti, 2011; Pawlowsky-Glahn, Egozcue and Tolosana-Delgado, 2015), as has dedicated user-friendly software (Boogaart and Tolosana-Delgado, 2013; Filzmoser et al., 2018; Greenacre, 2018; Palarea-Albaladejo and Martín-Fernández, 2015; Thió-Henestrosa and Martín-Fernández, 2005), although in many cases standard software can be used after transforming the data.

In compositional research problems, most of the basic statistical analysis tools are flawed unless they are re-expressed by means of logarithms of ratios as proposed in the so-called log-ratio CoDa methodology.

The appeal of log-ratios is that once they are computed, standard statistical methods can be used in many cases, as long as the relative character of the information is taken into account when interpreting the results. Since one component can only increase *in relative terms* if some other(s) decrease, the effects of components as explanatory in a regression model cannot be interpreted in isolation. The effect of increasing one component in relative terms unavoidably depends on which other components are reduced in its stead. We emphasise the phrase "in relative terms" according to a compositional research focus, because it could be the case that all components increase in absolute terms.

In this article we stress the importance of the notion that in ordinary-least-squares multiple regression models interpretation of a predictor is always subject to keeping all other predictors constant. In log-ratio terms, the effect of increasing one log-ratio is understood while keeping all other log-ratios constant. The fact that the same log-ratio can have different effects and interpretations depending on the manner in which the remaining log-ratios in the regression model are constructed is frequently overlooked by applied researchers. This notion may also make interpretation of log-ratios as explanatory variables differ from other statistical analyses, which is also often overlooked.

Many ways of constructing and interpreting log-ratios have been suggested in the literature, which often lead to the same predictions, residuals and goodness of fit of the model. Given this circumstance, it is difficult to provide arguments to choose among them, since "there seems to be little to distinguish between forms of comparable goodness of fit. Much discussion has turned on attempts to provide interpretations for the parameters" (Aitchison (1986), p285). In a sense, the alternative log-ratios do not lead to different models, but to different reparametrizations of one and the same model. Each reparametrization aims at one particular manner of interpreting the results. The aim of this article is to review some of the most common alternative parametrizations and highlight their implications regarding parameter interpretation, "keeping all other log-ratios in the model constant". To the best of our knowledge, some coincidences and similarities between the interpretations of these parametrizations are reported for the first time in this article, which will hopefully help researchers find their way in the crossroads of the many methods proposed in the literature. Some of the parametrizations involve rerunning the model more than once in order to shed additional light on the meaning of parameters.

The five particular parametrizations chosen in this article are aimed at easing interpretation and all have comparable and readily interpretable effect sizes, which are obscured in compositional analyses with more complicated alternatives, whose use prevents effect size interpretation from being a common practice in the applied literature (Müller et al., 2018).

The article starts with the first parametrizations, chronologically speaking (additive and centred log-ratios) and continues with some more recently proposed alternatives. Both statistical tests and effect sizes are interpreted and compared. An illustration using one of Aitchison's classic data sets follows. The last section concludes.

## 2 Basic form of the regression model with an explanatory composition. Additive log-ratios

Consider a *composition* $\mathbf{x}$, i.e. a vector in the positive orthant of $D$-dimensional real space carrying information about the relative importance of its components. For ease of formulation and illustration, in this article we consider $D = 4$ components closed to one without loss of generality:

$$\mathbf{x} = (x_1,\, x_2,\, x_3,\, x_4) \in R_+^4, \text{ with } x_j > 0, j = 1,2,3,4, \ \sum_{j=1}^{4} x_j = 1. \tag{1}$$

The most common CoDa approach is to represent $\mathbf{x}$ in terms of logarithms of ratios among its components (Aitchison, 1986; Egozcue et al., 2003). Log-ratios may, for instance, be computed among all possible pairs of components in the so-called *pairwise log-ratios* (Aitchison, 1986; Greenacre, 2019). In this article we follow Müller et al. (2018) in computing logarithms to base 2, which make for a simple interpretation. A unit increase in the logarithm to base 2 corresponds to a twofold increase in the original magnitude.

$$\log_2\left(\frac{x_j}{x_k}\right) = \log_2\left(x_j\right) - \log_2\left(x_k\right) \text{ , with } j < k, k = 2,3,4, \ j = 1,2,3. \tag{2}$$

A particularly interesting case of pairwise log-ratios is that of *additive log-ratios* (Aitchison, 1982), in which only $D-1$ pairwise log-ratios are computed with a common component in the denominator, for instance the last. This yields an invertible log-ratio covariance matrix and additive log-ratios can thus be directly used as predictors in an ordinary-least-squares regression model:

$$\log_2\left(\frac{x_j}{x_4}\right), \text{ with } j = 1,2,3. \tag{3}$$

The most useful and general expression of a log-ratio is the log-contrast (Aitchison, 1983; Aitchison and Bacon-Shone, 1984):

$$\sum_{j=1}^{4} \alpha_j \log_2(x_j), \quad \text{with } \sum_{j=1}^{4} \alpha_j = 0. \tag{4}$$

Log-contrasts led to the first formalization of a regression with a compositional explanatory variable (Aitchison and Bacon-Shone, 1984). The regression problem can be understood as obtaining the log-contrast which is maximally correlated with the dependent variable:

$$\begin{aligned} &y = \alpha_0 + \alpha_1 \log_2(x_1) + \alpha_2 \log_2(x_2) + \alpha_3 \log_2(x_3) + \alpha_4 \log_2(x_4) + \varepsilon, \\ &\text{with } \sum_{j=1}^{4} \alpha_j = 0, \end{aligned} \tag{5}$$

where $\epsilon$ follows the usual assumptions in the linear regression model. The zero-sum constraint of the coefficients in Eq. (5) reflects the fact that a component can only increase its relative importance if one or more of the others decrease. Geometrically, the zero sum constraint implies that the vector $[\alpha_1, \alpha_2, \alpha_3, \alpha_4]^T$ is orthogonal to the unit vector $[1,1,1,1]^T$ as required for a composition, which is key to the scale invariance property in CoDa (Egozcue and Pawlowsky-Glahn, 2019): multiplying the individual compositions by arbitrary positive constants will not modify the regression results. The constraint can be handled by running the regression model by ordinary least squares on the additive log-ratios (Aitchison and Bacon-Shone, 1984):

$$y = \beta_0 + \beta_1 \log_2\left(\frac{x_1}{x_4}\right) + \beta_2 \log_2\left(\frac{x_2}{x_4}\right) + \beta_3 \log_2\left(\frac{x_3}{x_4}\right) + \varepsilon. \tag{6}$$

All formulations presented in this article lead to the same coefficients when re-expressed according to Eq. (5). In the additive log-ratio case the re-expression and the fulfilment of the constraint are trivial:

$$y = \beta_0 + \beta_1 \log_2(x_1) + \beta_2 \log_2(x_2) + \beta_3 \log_2(x_3) + (-\beta_1 - \beta_2 - \beta_3) \log_2(x_4) + \varepsilon. \tag{7}$$

The $\alpha_4$ coefficient corresponding to $\log_2(x_4)$ can also be obtained by rerunning the model with a different denominator in the additive log-ratio transformation, for instance as the coefficient of $\log_2(x_4/x_3)$ in a model in which all log-ratios have $x_3$ in the denominator.

The interpretation is as follows. The expected value of the dependent variable increases when increasing the relative importance of components with positive $\alpha$ coefficients in Eq. (5), especially those with the largest coefficients in absolute value, at the expense of reducing that of components with negative $\alpha$ coefficients (especially those with the largest coefficients in absolute value). In compositional regression models, the interpretation can never be that of "increasing one component while keeping all other components constant" because this statement is, in relative terms, nonsensical.

Overall tests of all $D-1$ effects in Eq. (6) simultaneously (e.g., joint $F$ tests in linear regression) are invariant for all approaches described in this article. Since the composition is multivariate by nature, the joint test is normally the one with the greatest interest. Rejecting the null hypothesis means that the composition as a whole has an effect on the dependent variable.

Having said this, researchers sometimes like to test other more specific hypotheses. For this purpose, the proper interpretation of the coefficients of each log-ratio is crucial. Interpretation does change among the alternative approaches discussed in this article. In this section we interpret the coefficients of additive log-ratios. As in any multiple regression model by ordinary least squares, the effect of one log-ratio and its test is understood as the expected change in $y$ for a one-unit change of the log-ratio when the other log-ratios are held constant (Pindyck and Rubinfeld, 1976).

Accordingly, when using logarithms to base 2, $\beta_1$ is the effect of doubling the ratio between $x_1$ and $x_4$ *while keeping all other log-ratios constant*. Keeping the second log-ratio constant means that $x_2$ can only vary by the same factor as $x_4$. Keeping the third log-ratio constant means that $x_3$ can only vary by the same factor as $x_4$. The interpretation of $\beta_1$ is thus the change in the dependent variable expected value when the ratio between $x_1$ and each of components 2 to $D$ doubles. It is also the change in the dependent variable expected value when the ratio between $x_1$ and the geometric mean of all other components doubles, with the restriction that components 2 to $D$ vary by a common factor. All effect sizes are hence readily interpretable and comparable: the interpretation of $\beta_j$ is the change in the dependent variable expected value when the ratio between $x_j$ and each and every of the components $x_1,...,x_{j-1},x_{j+1},...,x_D$ doubles. If we consider the fact that in relative terms one component can only increase if other components decrease, statistically testing the $\beta_j$ parameter means testing if increasing the $x_j$ component at the expense of reducing all other components by a common factor has any impact on the dependent variable.

Table 1 shows an example of a fictitious population with $\beta_0 = 0$, $\beta_1 = 1$, $\beta_2 = 2$, and $\beta_3 = 3$ as in Eq. (6). It can be noted that, as compared to case 1, case 2 doubles the ratio of $x_1$ over each and every of the remaining components. As compared to case 1, case 2 increases the first log-ratio by one unit while keeping the remaining log-ratios constant.

As a result, as compared to case 1, case 2 increases the expected value E($y$) by $\beta_1 = 1$. The interested reader may compare cases 3 and 4 with case 1 to arrive at $\beta_2$ and $\beta_3$. The comparison between case 5 and case 1 leads to $\alpha_4$.

Regression effects can also be interpreted in terms of what in CoDa is known as the *perturbation* operator, which can be explained in brief as the product of two compositions, component-wise. Increases in log-ratios correspond to perturbations when expressed with respect to the original compositions. Therefore, Table 1 can also be interpreted in terms of perturbations. Cases 2 to 4 in the table correspond to the perturbation of the $[x_1, x_2, x_3, x_4]$ composition when increasing each log-ratio by one unit. For instance, increasing $\log_2\left(\frac{x_1}{x_4}\right)$ by one unit while keeping all other log-ratios constant is equivalent to perturbing the original composition with $[0.4, 0.2, 0.2, 0.2]$. The product of $[0.4, 0.2, 0.2, 0.2]$ and $[0.25, 0.25, 0.25, 0.25]$ yields $[0.1, 0.05, 0.05, 0.05]$ which is closed back to a unit sum as $[0.4, 0.2, 0.2, 0.2]$. The *inverse log-ratio transformation* is the manner in which log-ratios can be expressed back as the original composition. It should be clear that these perturbations are nothing other than the inverse log-ratio transformations of vectors $[1, 0, 0]$, $[0, 1, 0]$ and $[0, 0, 1]$. Indeed, using logarithms to base 2, the inverse additive log-ratio transformation of $[1, 0, 0]$ is $[2^1, 2^0, 2^0, 2^0]$ which, after closing to unit sum, equals $[0.4, 0.2, 0.2, 0.2]$.

***Table 1***: *Fictitious population with* $\beta_0 = 0, \beta_1 = 1, \beta_2 = 2, \beta_3 = 3$. *Additive log-ratios.*

| *Case* | $x_1$ | $x_2$ | $x_3$ | $x_4$ | $\log_2\left(\frac{x_1}{x_4}\right)$ | $\log_2\left(\frac{x_2}{x_4}\right)$ | $\log_2\left(\frac{x_3}{x_4}\right)$ | E($y$) |
|---|---|---|---|---|---|---|---|---|
| 1 | 0.250 | 0.250 | 0.250 | 0.250 | 0 | 0 | 0 | 0 |
| 2 | 0.400 | 0.200 | 0.200 | 0.200 | 1 | 0 | 0 | 1 |
| 3 | 0.200 | 0.400 | 0.200 | 0.200 | 0 | 1 | 0 | 2 |
| 4 | 0.200 | 0.200 | 0.400 | 0.200 | 0 | 0 | 1 | 3 |
| 5 | 0.200 | 0.200 | 0.200 | 0.400 | −1 | −1 | −1 | −6 |

It must be noted that even if the construction of the log-ratio $\log_2(x_j/x_4)$ suggests increasing $x_j$ in relative terms to only $x_4$, this does not correspond to its interpretation when the composition is explanatory, because control of the other log-ratios is a key issue.

## 3 Regression model with explanatory centred log-ratios

Log-ratios are often computed between each component and the geometric mean of all components including itself, in the so-called *centred log-ratio* (Aitchison, 1983):

$$\log_2\left(\frac{x_j}{\sqrt[4]{x_1 x_2 x_3 x_4}}\right), \text{ with } j = 1, 2, 3, 4. \tag{8}$$

In order to prevent perfect collinearity, one centred log-ratio must be dropped from the regression equation. This is by no means a nuisance, as often argued, but is the key to the proper parameter interpretation, as shown below. Without loss of generality, if we leave out the last centred log-ratio, the model formulation is:

$$y=\beta_0+\beta_1\log_2\left(\frac{x_1}{\sqrt[4]{x_1x_2x_3x_4}}\right)+\beta_2\log_2\left(\frac{x_2}{\sqrt[4]{x_1x_2x_3x_4}}\right)+\beta_3\log_2\left(\frac{x_3}{\sqrt[4]{x_1x_2x_3x_4}}\right)+\varepsilon. \tag{9}$$

By expressing Eq. (9) as a the log-contrast in Eq. (5) we obtain:

$$y=\beta_0+\left(\beta_1-\frac{1}{4}\sum_{j=1}^{3}\beta_j\right)\log_2(x_1)+\left(\beta_2-\frac{1}{4}\sum_{j=1}^{3}\beta_j\right)\log_2(x_2)+$$

$$\left(\beta_3-\frac{1}{4}\sum_{j=1}^{3}\beta_j\right)\log_2(x_3)+\left(-\frac{1}{4}\sum_{j=1}^{3}\beta_j\right)\log_2(x_4)+\varepsilon. \tag{10}$$

Univariate tests referring to each particular log-ratio are interpreted as follows. Since all four centred log-ratios in Eq. (8) add-up to zero, increasing a given centred log-ratio while keeping the remaining two log-ratios in the equation constant means increasing the given centred log-ratio while decreasing the omitted centred log-ratio by the same amount. Individual coefficients and their tests thus show the existence of significant trade-offs between pairs of components. A positive significant $\beta_j$ coefficient means that increasing component $x_j$ at the expense of reducing component $x_4$ has a significant positive effect on the dependent variable. If we use the logarithm to base 2, $\beta_j$ is interpreted as the expected change in the dependent variable when the ratio between $x_j$ and $x_4$ increases fourfold. In this manner effect sizes in the model are once more readily interpretable and comparable. Table 2 has an example of a fictitious population with $\beta_0=0$, $\beta_1=1$, $\beta_2=2$, and $\beta_3=-1$ as in Eq. (9). For instance, as compared to case 1, case 2 increases the first log-ratio by one unit while keeping the remaining log-ratios constant and shows a fourfold increase in the ratio between $x_1$ and $x_4$. Compared to case 1, the ratio between $x_1$ and any of the remaining components ($x_2$ and $x_3$) is doubled, while the ratio between $x_4$ and any of the remaining components ($x_2$ and $x_3$) is halved. It must be noted that the omitted log-ratio $\log_2\left(\frac{x_4}{\sqrt[4]{x_1x_2x_3x_4}}\right)$ is equal to $-1$.

***Table 2***: *Fictitious population with $\beta_0=0,\beta_1=1,\beta_2=2,\beta_3=-1$. Centred log-ratios, where the one with $x_4$ in the numerator has been dropped.*

| *Case* | $x_1$ | $x_2$ | $x_3$ | $x_4$ | $\log_2\left(\frac{x_1}{\sqrt[4]{x_1x_2x_3x_4}}\right)$ | $\log_2\left(\frac{x_2}{\sqrt[4]{x_1x_2x_3x_4}}\right)$ | $\log_2\left(\frac{x_3}{\sqrt[4]{x_1x_2x_3x_4}}\right)$ | E($y$) |
|---|---|---|---|---|---|---|---|---|
| 1 | 0.250 | 0.250 | 0.250 | 0.250 | 0 | 0 | 0 | 0 |
| 2 | 0.444 | 0.222 | 0.222 | 0.111 | 1 | 0 | 0 | 1 |
| 3 | 0.222 | 0.444 | 0.222 | 0.111 | 0 | 1 | 0 | 2 |
| 4 | 0.222 | 0.222 | 0.444 | 0.111 | 0 | 0 | 1 | −1 |

In order to get tests and estimates for all possible pairwise trade-offs, the model can be rerun $D$ times by dropping each time a different centred log-ratio.

It must be noted once more that even if the construction of the log-ratio $\log_2\left(\frac{x_j}{\sqrt[4]{x_1x_2x_3x_4}}\right)$ suggests increasing $x_j$ in relative terms to all components, this does not correspond to its interpretation when the composition is explanatory, because control of the other log-ratios is a key issue.

## 4 Regression model with explanatory pivot coordinates

Egozcue et al. (2003) were the first to advocate for an orthonormal basis to compute the log-ratio transformation. The advantages of this approach in many statistical analyses can be found in Pawlowsky-Glahn et al. (2015). This recommendation was translated to models with explanatory compositions by Tolosana-Delgado and Boogaart (2011) in the form of *balance coordinates* (Egozcue and Pawlowsky-Glahn, 2005). In short, balance coordinates are *scaled log-ratios* of the geometric means of two groups of components, chosen in such a way that the basis is orthonormal.

Within balance coordinates, one particular form (Egozcue et al., 2003; Fišerová and Hron, 2011; Hron, Filzmoser and Thompson, 2012) which later became known as *pivot coordinates* (Filzmoser et al., 2018), makes it possible to interpret the effect of increasing one component at the expense of decreasing all others by a common factor and has gained widespread acceptance, partly due to the unawareness that the original formulation as additive log-ratios by Aitchison and Bacon-Shone (1984) is interpreted in the same manner up to a scaling constant when used as explanatory (Coenders, 2019).

In order to provide an easily interpretable and comparable measure of effect size, Müller et al. (2018) wisely changed the requirement of orthonormality of the basis to mere orthogonality by removing scaling constants from pivot coordinates, unaware that this resulted in the same estimates and test statistics as the additive log-ratio representation. This approach was first referred to as *orthogonal coordinates for compositional regression* (Müller et al., 2018). Henceforth we refer to them as *simplified pivots*.

The first coordinate under the simplified pivot approach is the log-ratio of the first component over the geometric mean of all other components, the second is the log-ratio of the second component over the geometric mean of components 3 to $D$, the third is the log-ratio of the third component over the geometric mean of components 4 to $D$, and so forth. Constructed as just described, the following log-ratios make it possible to interpret the first log-ratio, which is the one to be called *pivot*, as the effect of increasing the first component while reducing all others by a common factor (Hron et al., 2012; Müller et al., 2018):

$$y = \beta_0 + \beta_1 \log_2\left(\frac{x_1}{\sqrt[3]{x_2x_3x_4}}\right) + \beta_2 \log_2\left(\frac{x_2}{\sqrt[2]{x_3x_4}}\right) + \beta_3 \log_2\left(\frac{x_3}{x_4}\right) + \varepsilon. \quad (11)$$

The model can be rerun $D$ times by permuting the components so that each time one different component plays the role of the first, which is in the numerator of the first log-ratio. The order of all other components is irrelevant. Each run provides one of the $\alpha$ coefficients in Eq. (5).

Table 3 shows an example of a fictitious population with $\beta_0 = 0$, $\beta_1 = 1$, $\beta_2 = 2$, and $\beta_3 = 3$ as in Eq. (11). The reader will note that keeping the second and third log-ratios constant forces all components in the denominator of the first log-ratio to change by a common factor. Thus, as compared to case 1, case 2 doubles the ratio of $x_1$ over each and every of the remaining components, exactly as in Table 1. Also as compared to case 1, case 2 increases the first log-ratio by one unit while keeping the remaining log-ratios constant. Cases 3 and 4 and coefficients $\beta_2$ and $\beta_3$ are usually not interpreted in the pivot coordinate case.

***Table 3***: *Fictitious population with* $\beta_0 = 0, \beta_1 = 1, \beta_2 = 2, \beta_3 = 3$. *Simplified pivots.*

| *Case* | $x_1$ | $x_2$ | $x_3$ | $x_4$ | $\log_2\left(\frac{x_1}{\sqrt[3]{x_2 x_3 x_4}}\right)$ | $\log_2\left(\frac{x_2}{\sqrt[2]{x_3 x_4}}\right)$ | $\log_2\left(\frac{x_3}{x_4}\right)$ | E($y$) |
|---|---|---|---|---|---|---|---|---|
| 1 | 0.250 | 0.250 | 0.250 | 0.250 | 0 | 0 | 0 | 0 |
| 2 | 0.400 | 0.200 | 0.200 | 0.200 | 1 | 0 | 0 | 1 |
| 3 | 0.240 | 0.380 | 0.190 | 0.190 | 0 | 1 | 0 | 2 |
| 4 | 0.243 | 0.243 | 0.343 | 0.172 | 0 | 0 | 1 | 3 |

Since only the $\beta_1$ coefficient is interpreted in each of the $D$ model runs, sometimes researchers compile a table including only these, which can give the misleading impression that there is only one regression model with $D$ log-ratios while there actually are $D$ regression models, each with $D-1$ log-ratios. We do not discuss further the estimates and tests and their interpretation because they are identical to the additive log-ratio case, albeit in the simplified pivot case, interpretation is more intuitive in accordance with the way in which the log-ratio is constructed.

## 5 Regression model with other explanatory orthogonal coordinates

Besides pivot coordinates, any balance coordinates can be re-expressed as orthogonal coordinates for compositional regression (Müller et al., 2018) by just dropping the scaling constants. They are thus just the logarithm of the geometric means of two groups of components, one in the numerator and one in the denominator taking care that the basis is orthogonal. As ordinary balance coordinates, they can be formed from a *sequential binary partition* of the components (Egozcue and Pawlowsky-Glahn, 2005). There are potentially many ways in which components can be partitioned, and the choice can be tailored to the research objectives. We provide only an example: we firstly partition the whole composition into the group of components $x_1$ and $x_2$ on the one hand and the group $x_3$ and $x_4$ on the other, we secondly partition the first group into its two single components, and thirdly we do likewise with the second group, according to the rows of the following sign matrix, in which positive signs indicate the numerator of the log-ratio and negative signs the denominator:

$$\begin{array}{cccc} x_1 & x_2 & x_3 & x_4 \\ +1 & +1 & -1 & -1 \\ +1 & -1 & 0 & 0 \\ 0 & 0 & +1 & -1 \end{array}. \tag{12}$$

We get the following reparametrization:

$$y = \beta_0 + \beta_1 \log_2\left(\frac{\sqrt[2]{x_1 x_2}}{\sqrt[2]{x_3 x_4}}\right) + \beta_2 \log_2\left(\frac{x_1}{x_2}\right) + \beta_3 \log_2\left(\frac{x_3}{x_4}\right) + \varepsilon. \tag{13}$$

Since it is feasible to compute potentially many sets of orthogonal coordinates by partitioning the components in different ways, the interpretation has to be tailored to the particular log-ratios. If we consider what it means to keep the second and third log-ratios constant while interpreting the first one, the estimates and tests of $\beta_1$ have to be interpreted as the effect of increasing $x_1$ and $x_2$ by a common factor and reducing $x_3$ and $x_4$ by a common factor in such a way that the ratio of the geometric means of the first pair over the second doubles (assuming we use the logarithm to base 2). A sequential binary partition chosen by the researcher as in Eq. (12) makes it possible to test the effect of jointly increasing *any subset of components* by a common factor while decreasing *any other subset of components* by a common factor. If we consider what it means to keep the first and third log-ratio constant while interpreting the second log-ratio, the estimates and tests of $\beta_2$ have to be interpreted as the effect of doubling the $x_1$ to $x_2$ ratio without modifying the relative importance of $x_3$ to $x_4$, nor the relative importance of $x_3$ and $x_4$ to $x_1$ and $x_2$ in geometric mean terms, in the same way as in the centred log-ratio case. The actual estimate is half of that obtained with the centred log-ratio and the test result is identical. The reader will note that, by coincidence, the formulation of the last log-ratio in Eq. (13) coincides with the additive log-ratio case in Eq. (6), but not its interpretation.

Table 4 has an example of a fictitious population with $\beta_0 = 0$, $\beta_1 = 1$, $\beta_2 = 2$, and $\beta_3 = -2$ as in Eq. (13). For instance, as compared to case 1, case 2 increases the first log-ratio by one unit while keeping the remaining log-ratios constant and shows a twofold increase in the ratio between $x_1$ and $x_2$ on the one hand and $x_3$ and $x_4$ on the other.

Summing up, orthogonal coordinates for compositional regression have the attractive property that effects can always be interpreted as increasing the components in the numerator by a common factor while decreasing those in the denominator by a common factor in such a way that the ratio has a twofold increase. The perturbation $[0.333, 0.333, 0.167.0.167]$ in the second row of Table 4 associated to the $\log_2\left(\frac{\sqrt[2]{x_1 x_2}}{\sqrt[2]{x_3 x_4}}\right)$ log-ratio is a good example. In the orthogonal coordinate case, interpretation is intuitive in accordance with the way in which the log-ratio is constructed.

***Table 4***: *Fictitious population with* $\beta_0 = 0, \beta_1 = 1, \beta_2 = 2, \beta_3 = -2$. *Orthogonal coordinates for compositional regression.*

| *Case* | $x_1$ | $x_2$ | $x_3$ | $x_4$ | $\log_2\left(\frac{\sqrt[2]{x_1x_2}}{\sqrt[2]{x_3x_4}}\right)$ | $\log_2\left(\frac{x_1}{x_2}\right)$ | $\log_2\left(\frac{x_3}{x_4}\right)$ | E($y$) |
|---|---|---|---|---|---|---|---|---|
| 1 | 0.250 | 0.250 | 0.250 | 0.250 | 0 | 0 | 0 | 0 |
| 2 | 0.333 | 0.333 | 0.167 | 0.167 | 1 | 0 | 0 | 1 |
| 3 | 0.343 | 0.172 | 0.243 | 0.243 | 0 | 1 | 0 | 2 |
| 4 | 0.243 | 0.243 | 0.343 | 0.172 | 0 | 0 | 1 | −2 |

Of course, the original balance coordinates which include scaling constants are equivalent up to a change in scale, which preserves the statistical test results but makes effect sizes less readily comparable.

## 6 Regression model with explanatory pairwise log-ratios

Greenacre (2019) suggested a general approach to selecting $D-1$ pairwise log-ratios, which, when introduced as explanatory, provide yet another flexible way of testing hypotheses that can be tailored to the research objectives. It boils down to taking care that each component participates in at least one log-ratio and that exactly $D-1$ log-ratios are computed. This results in an acyclic connected graph in which the $D$ components act as nodes and the $D-1$ log-ratios as edges (Greenacre, 2019). Once more, this makes for a very high number of possible reparametrizations. As in the section above, we present just one example. The reader will note that the formulation of the first log-ratio coincides with the additive log-ratio case in Eq. (6), but not its interpretation.

$$y = \beta_0 + \beta_1 \log_2\left(\frac{x_1}{x_4}\right) + \beta_2 \log_2\left(\frac{x_2}{x_1}\right) + \beta_3 \log_2\left(\frac{x_3}{x_4}\right) + \varepsilon. \qquad (14)$$

The log-ratios in Eq. (14) correspond to the graph in Figure 1.

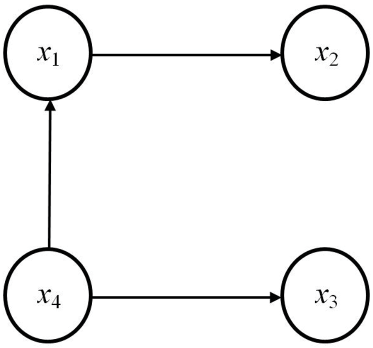

***Figure 1***: *Acyclic connected graph representing the pairwise log-ratios. Arrows point from the denominator to the numerator.*

When interpreting each log-ratio the researcher will have to be well aware of what the model is controlling for. If we use the logarithm to base 2, $\beta_1$ is interpreted as the effect of doubling the ratio of $x_1$ over $x_4$. However, keeping the second log-ratio constant means that $x_2$ can only increase by the same factor as $x_1$, and keeping the third log-ratio constant means that $x_3$ can only decrease by the same factor as $x_4$. The graph in Figure *1* also shows $x_2$ to be connected to $x_1$, and $x_3$ to $x_4$. Thus the estimates and tests of $\beta_1$ have to be interpreted as the effect of multiplying $x_1$ and $x_2$ by a common factor and $x_3$ and $x_4$ by a common factor in such a way that the ratio of the first pair over the second doubles. The interpretation is thus not the same as in the additive log-ratio case, even if the formulation of the log-ratio is the same as in Eq. (6).

Rearranging the components in the pairwise log-ratios makes it possible to test the effect of jointly increasing any subset of components by a common factor while decreasing *all the remaining components by a common factor*. The remaining log-ratios and the acyclic connected graph inform the researcher of which other components are linked to the numerator and which to the denominator of the log-ratio which is being interpreted, for which purpose great care has to be exerted. When used as explanatory, pairwise log-ratios are thus more closely related to orthogonal coordinates for compositional regression than previously thought, although less flexible, because orthogonal coordinates make it possible to leave certain components out of both the denominator and the numerator for interpretation.

Because of the way in which the pairwise log-ratios are computed in this particular example, the reader can apply the reasoning above to find out that the interpretation of $\beta_2$ and $\beta_3$ is the same as in the additive log-ratio case, and also corresponds to two particular simplified pivots. For instance, when interpreting $\beta_2$, keeping the first and third log-ratios constant implies that $x_4$ and $x_3$ vary by the same factor as $x_1$, respectively, while no component varies by the same factor as $x_2$. The graph in Figure *1* also shows $x_4$ and $x_3$ to be connected to $x_1$. $\beta_2$ thus refers to doubling the ratio of $x_2$ over all other components assuming that they decrease by a common factor.

Table 5 shows an example of a fictitious population with $\beta_0 = 0$, $\beta_1 = 2$, $\beta_2 = 1$, and $\beta_3 = -1$ as in Eq. (14), which illustrates the interpretations above.

***Table 5***: *Fictitious population with* $\beta_0 = 0, \beta_1 = 2, \beta_2 = 1, \beta_3 = -1$. *Pairwise log-ratios.*

| *Case* | $x_1$ | $x_2$ | $x_3$ | $x_4$ | $\log_2\left(\frac{x_1}{x_4}\right)$ | $\log_2\left(\frac{x_2}{x_1}\right)$ | $\log_2\left(\frac{x_3}{x_4}\right)$ | E($y$) |
|---|---|---|---|---|---|---|---|---|
| 1 | 0.250 | 0.250 | 0.250 | 0.250 | 0 | 0 | 0 | 0 |
| 2 | 0.333 | 0.333 | 0.167 | 0.167 | 1 | 0 | 0 | 2 |
| 3 | 0.200 | 0.400 | 0.200 | 0.200 | 0 | 1 | 0 | 1 |
| 4 | 0.200 | 0.200 | 0.400 | 0.200 | 0 | 0 | 1 | $-1$ |

It must be noted again that even if the construction of the log-ratio $\log_2(x_j/x_k)$ suggests increasing $x_j$ in relative terms to only $x_k$, this does not correspond to its interpretation when the composition is explanatory, because control of the other log-ratios is a key issue.

It must be reminded that cases 2 to 4 in all Tables 1 to 5 can also be interpreted in terms of the perturbation of the $[x_1, x_2, x_3, x_4]$ composition when increasing each log-ratio by one unit while keeping all other log-ratios constant. In all cases, the perturbations are obtained as the inverse transformations of vectors $[1,0,0]$, $[0,1,0]$ and $[0,0,1]$ according to the given log-ratio transformation.

## 7 Illustration

As an illustration we use one of the original simulated data sets provided by Aitchison (1986), called *Bayesite*, which is freely available in the R library compositions (Boogaart and Tolosana-Delgado, 2013).

In the development of bayesite, a new fibreboard, experiments were conducted to obtain some insight into the nature of the relationship of its permeability (measured in microdarcies) to the mix of its four ingredients ($n = 21$):

- Short fibres ($x_1$).
- Medium fibres ($x_2$).
- Long fibres ($x_3$).
- Binder ($x_4$)

All model parametrizations have an intercept term equal to 317.302, a residual standard error equal to 46.61 on 17 degrees of freedom, multiple R-squared equal to 0.419, adjusted R-squared equal to 0.316, and a significant joint $F$ statistic (4.078 on 3 and 17 degrees of freedom, p-value = 0.024), telling that the mix as a whole has an impact on permeability.

Table 6 shows the coefficients. Those in italics are either redundant or not needed for interpretation. The most correlated log-contrast with permeability is also the same for all parametrizations:

$$19.414\log_2(x_1) + 27.406\log_2(x_2) - 25.953\log_2(x_3) - 20.866\log_2(x_4). \qquad (15)$$

This means that permeability increases together with increases of $x_1$ and $x_2$ coupled with decreases in $x_3$ and $x_4$, in terms of relative importance, $x_2$ and $x_3$ having a greater impact than $x_1$ and $x_4$.

The additive log-ratio results tell that increasing the relative importance of $x_2$ at the expense of reducing all other components by a common factor in such a way that the ratio of $x_2$ over any other component doubles, leads to an expected increase of 27.406 microdarcies in permeability, which is statistically significant at the 0.05 level. The results also tell that relatively increasing $x_3$ at the expense of reducing all other components by a common factor in such a way that the ratio of $x_3$ over any other component doubles, leads to a significant expected decrease of 25.953 microdarcies.

***Table 6***: *Estimates and tests in four alternative reparametrizations. Redundant or not needed effects in italics.*

| Additive log-ratios: | Estimate | Std. Error | t-value | p-value |
|---|---|---|---|---|
| Denominator $x_4$ | | | | |
| Numerator $x_1$ | 19.414 | 11.560 | 1.679 | 0.111 |
| Numerator $x_2$ | 27.406 | 11.560 | 2.371 | 0.030 |
| Numerator $x_3$ | −25.953 | 11.560 | −2.245 | 0.038 |
| Denominator $x_3$ | | | | |
| *Numerator* $x_1$ | *19.414* | *11.560* | *1.679* | *0.111* |
| *Numerator* $x_2$ | *27.406* | *11.560* | *2.371* | *0.030* |
| Numerator $x_4$ | −20.866 | 16.229 | −1.286 | 0.216 |
| Centred log-ratios: | | | | |
| Log-ratio with $x_4$ numerator omitted | | | | |
| Numerator $x_1$ | 40.281 | 23.929 | 1.683 | 0.111 |
| Numerator $x_2$ | 48.272 | 23.929 | 2.017 | 0.060 |
| Numerator $x_3$ | −5.087 | 23.929 | −0.213 | 0.834 |
| Log-ratio with $x_3$ numerator omitted | | | | |
| Numerator $x_1$ | 45.368 | 17.694 | 2.564 | 0.020 |
| Numerator $x_2$ | 53.360 | 17.694 | 3.016 | 0.008 |
| *Numerator* $x_4$ | *5.087* | *23.929* | *0.213* | *0.834* |
| Log-ratio with $x_2$ numerator omitted | | | | |
| Numerator $x_1$ | −7.991 | 17.694 | −0.452 | 0.657 |
| *Numerator* $x_3$ | *−53.360* | *17.694* | *−3.016* | *0.008* |
| *Numerator* $x_4$ | *−48.272* | *23.929* | *−2.017* | *0.060* |
| Log-ratio with $x_1$ numerator omitted | | | | |
| *Numerator* $x_2$ | *7.991* | *17.694* | *0.452* | *0.657* |
| *Numerator* $x_3$ | *−45.368* | *17.694* | *−2.564* | *0.020* |
| *Numerator* $x_4$ | *−40.281* | *23.929* | *−1.683* | *0.111* |
| Simplified pivots: | | | | |
| $x_1$ in the first place | | | | |
| Pivot | 19.414 | 11.560 | 1.679 | 0.111 |
| *Second log-ratio* | *33.877* | *11.541* | *2.935* | *0.009* |
| *Third log-ratio* | *−2.544* | *11.964* | *−0.213* | *0.834* |
| $x_2$ in the first place | | | | |
| Pivot | 27.406 | 11.560 | 2.371 | 0.030 |
| *Second log-ratio* | *28.550* | *11.541* | *2.474* | *0.024* |
| *Third log-ratio* | *−2.544* | *11.964* | *−0.213* | *0.834* |
| $x_3$ in the first place | | | | |
| Pivot | −25.953 | 11.560 | −2.245 | 0.038 |
| *Second log-ratio* | *10.763* | *11.541* | *0.933* | *0.364* |
| *Third log-ratio* | *24.136* | *11.964* | *2.017* | *0.060* |
| $x_4$ in the first place | | | | |
| Pivot | −20.866 | 16.229 | −1.286 | 0.216 |
| *Second log-ratio* | *12.459* | *10.216* | *1.220* | *0.239* |
| *Third log-ratio* | *26.680* | *8.847* | *3.016* | *0.008* |
| Other orthogonal coordinates (example) | | | | |
| $\log_2\left(\sqrt{x_1x_2}/\sqrt{x_3x_4}\right)$ | 46.820 | 14.880 | 3.147 | 0.006 |
| $\log_2(x_1/x_2)$ | −3.996 | 8.847 | −0.452 | 0.657 |
| $\log_2(x_3/x_4)$ | −2.544 | 11.964 | −0.213 | 0.834 |
| Pairwise log-ratios (example) | | | | |
| $\log_2(x_1/x_4)$ | 46.820 | 14.880 | 3.147 | 0.006 |
| $\log_2(x_2/x_1)$ | 27.406 | 11.560 | 2.371 | 0.030 |
| $\log_2(x_3/x_4)$ | −25.953 | 11.560 | −2.245 | 0.038 |

The centred log-ratio formulation shows that increasing $x_1$ at the expense of reducing $x_3$ and increasing $x_2$ at the expense of reducing $x_3$ both lead to a significant increase in permeability. Doubling $x_1$ at the expense of halving $x_3$ (i.e., multiplying their ratio by four) leads to a 45.368 microdarcy increase in expected permeability, while multiplying the ratio between $x_2$ and $x_3$ by four leads to a 53.360 increase in expected permeability.

The results with simplified pivots are identical numerically and interpreted in the same way as those with additive log-ratios.

In this particular example, the second and third orthogonal coordinates are tradeoffs between pairs of components and are thus related to the results of a centred log-ratio (estimates are halved but test statistics are identical). For instance, the test statistic for the coordinate $\log_2(x_1/x_2)$ is equivalent to the $x_1$ statistic in the centred log-ratio formulation with $x_2$ omitted.

Also in this particular example, the second and third pairwise log-ratios provide the same result as in the additive log-ratio case. Researchers need to carefully tailor interpretation to the particular log-ratios chosen, especially in the pairwise case. For instance, keeping the first and third pairwise log-ratios constant while increasing the second implies increasing the ratio of $x_2$ over all other components by the same factor.

Finally, the results of the first log-ratio both in the particular pairwise log-ratio example and the particular orthogonal coordinate example we have chosen, show that the effect of multiplying $x_1$ and $x_2$ by a common factor and $x_3$ and $x_4$ by another common factor in such a way that the ratio of the geometric mean of the first pair over the second doubles is significant, and amounts to 46.820, in terms of expected permeability in microdarcies.

## 8 Discussion

One attractive feature of CoDa is that once the raw composition has been transformed into log-ratios, classical statistical techniques for unbounded data can, in many cases, be applied in the usual way, and even with standard software. Log-ratio transformations thus constitute the easy way out in compositional problems. This includes models in which the composition is the explanatory variable. The applied researcher can concentrate his or her efforts in interpreting the results taking the compositional nature of the data and the research questions into account: what does increase at the expense of decreasing what? Along these lines, some quick and useful highlights to recap the article are:

- All alternatives considered in this article are reparametrizations of the same model. In a sense, none can be worse or better than any other as long as the parametrization provides answers to the researcher's questions and, above all, is interpreted correctly. If the researcher wants to interpret the results from more than one perspective or to test more than one type of hypotheses, he or she can use more than one parametrization.

- When used as explanatory variables, additive log-ratios are not interpreted as increasing a component at the expense of reducing the last component, as their formulation suggests, but as increasing a component at the expense of reducing *all other components*.
- When used as explanatory variables, centred log-ratios are not interpreted as increasing a component at the expense of reducing all other components, as their formulation suggests, but as increasing a component at the expense of reducing *the component whose log-ratio is omitted*.
- When used as explanatory variables, simplified pivot coordinates are equivalent to additive log-ratios.
- Orthogonal coordinates can be tailored to testing particular hypotheses of interest related to increasing any subset of components at the expense of reducing any other subset. Moreover, the interpretation of the regression coefficients is intuitive following the formulation of the corresponding log-ratios.
- When used as explanatory variables, pairwise log-ratios are not interpreted as increasing a component at the expense of reducing another component, as their formulation suggests, but as a tailored tool to interpret the effect of increasing a subset of components at the expense of reducing all the remaining components. Proper interpretation requires exerting great care.
- It often pays to embed theoretical knowledge or research questions into (possibly more than one) parametrizations of the model.
- As in any multiple regression, the full formulation of the model has much to tell about the log-ratio whose effect is being interpreted.
- Using logarithms to base 2 and removing scaling constants enhances interpretability and provides comparable effect size estimates.
- Expressing the effects of the log-ratios as the effects of the corresponding perturbations may help clarify their interpretation under all approaches, and even more so when tailoring orthogonal coordinates and pairwise log-ratios to the research objectives. The effect of the first log-ratio in the regression equation is that of perturbing the composition with the corresponding inverse log-ratio transformation of vector $[1,0,0,0,...,0]$, the second log-ratio refers to perturbing the composition with the inverse of vector $[0,1,0,0,...,0]$, and so on.

Great care must be taken if using the test results for simplifying the model (Pawlowsky-Glahn et al., 2015). The significance of one log-ratio depends both on the components present in the analysis and on the remaining $D-2$ log-ratios, which jointly frame the interpretation as the significant effect of relatively increasing what and how at the expense of relatively decreasing what and how. If we put it otherwise, dropping log-ratios changes the interpretation and estimates of whatever is left in the model. For instance, if we drop the second and third log-ratios in the orthogonal coordinate case in Eq. (13), then the coefficient of the first log ratio loses its original sharpness and shifts its interpretation into merely increasing the ratio of the product $x_1x_2$ over the product $x_3x_4$, without knowing if increases and decreases are by a common factor. This is so because the perturbation can no longer be computed from the inverse transformation. If we put it yet otherwise, all methods are interpreted with respect to the given set of components in $\mathbf{x}$. For instance, if we drop $x_1$ in the additive log-ratio case in Eq. (6), then the interpretation of the coefficient of $\log_2(x_2/x_4)$ is the outcome of increasing $x_2$ while decreasing only $x_3$ and $x_4$ by a common factor. Besides that, different parametrizations will unavoidably suggest different simplifications, and the simplified models will no longer be equivalent. In the bayesite example, the pivot and additive log-ratio approaches suggest that $x_1$ may be dropped from the composition whereas the centred log-ratio approach reveals a significant trade-off between $x_1$ and $x_3$. The centred log-ratio approach suggests to drop $x_4$ while in the pairwise log-ratio approach all the coefficients are significant. The easy way out is not to simplify the model at all. Of course, if the research is carried out for predictive or exploratory purposes rather than for theory building or theory testing, then simplifying the model can be the wise path to follow (see below) and parameter interpretation may not be essential.

We have not dealt with the diagnostic tools used in linear regression models with a compositional predictor because they are the same as in the general linear regression case, according to the distributional assumptions for the $\epsilon$ disturbance term, for instance the normal distribution. Any of the parametrizations can be obtained from any other parametrization by linear transformations. Since the regression model is affine equivariant, this implies that all parametrizations lead to the same goodness of fit, residuals, predicted values, and even leverage values and Cook's distances (Filzmoser et al., 2018). Conversely, the statistical distribution of the compositional variables plays no specific role. For this reason, orthogonality or isometry do not constitute requirements for using compositions as explanatory variables.

Having said this, orthogonal isometric log-ratios, among which balance coordinates constitute a common example, have very desirable properties in other compositional analyses, and can be blindly applied with virtually any statistical method. In the explanatory role, any orthogonal coordinates, isometric or not, also have the attractive property that effects can always be interpreted as increasing the components in the numerator by a common factor while decreasing those in the denominator by a common factor. Both advantages have no doubt contributed to their widespread use.

The extension from a linear model to a generalized linear model is straightforward (Coenders, Martín-Fernández and Ferrer-Rosell, 2017). For instance, if the dependent variable is a count, a *Poisson regression* can be specified, or if the dependent variable is ordinal or binary, an ordered or a binary *logit model* can be specified. Interpretation would then refer to the log expected count, to the logit, or to the appropriate expression in each case, taking the link function of the generalized linear model into account.

Adding non-compositional predictors in the same model can also be done in a straightforward manner (Coenders et al., 2017) and nested models can be used to assess the predictive power of the compositional versus non-compositional predictors. The results of the non-compositional predictors are invariant under any of the parametrizations of the composition presented in this article. The interpretation of the compositional predictors is the same as outlined in this article "keeping the non-compositional predictors constant". A very interesting particular case is including the total as predictor, which Coenders et al. (2017) recommend doing when the composition does not have a constant sum.

This article is by no means comprehensive. We have purposely selected only the simplest parametrizations with comparable effect sizes and leading to the same predictions. There are other ways to introduce a composition as explanatory in a regression model. A first group of methods (stability-based model selection, stepwise selection of the pairwise log-ratios with the highest explanatory power, spike-and-slab lasso regression modelling, principal balances, selection of the balance coordinate with highest explanatory power, and compositional principal component analysis, among others) always simplify the model, each in its own way, and thus lead to different predictions, do not control for all possible components or all possible log-ratios, and modify the interpretation. The interested reader may resort to the original sources (Combettes and Müller, 2019; Greenacre, 2019; Lin et al., 2014; Louzada, Shimizu and Suzuki, 2019; Martín-Fernández et al., 2018; Quinn and Erb, 2020; Rivera-Pinto et al., 2018; Solans et al., 2019). These data-driven approaches are especially useful when the number of components is very large, sometimes even larger than the sample size, when the model is built with predictive purposes, or when theory is weak and the researcher prefers to embrace a data mining perspective. A second group of methods does not imply simplifying the model. Among them we highlight interpreting the effects of balance coordinates, which up to a scaling constant are equivalent to those of orthogonal coordinates for compositional regression (Pawlowsky-Glahn et al., 2015), comparing predictions with different composition values (Dumuid et al., 2019), converting estimates into a gradient (Tolosana-Delgado and Boogaart, 2011), and converting estimates into elasticities (Morais, Thomas-Agnan and Simioni, 2018).

Having said this, we hope that by focusing on the most simple alternatives and on the comparative interpretation of their effect sizes and tests, we make it easier for researchers to draw fruitful, precise and clear conclusions about the influence of a composition on a dependent variable. Especially, a focus on effect sizes is as of now lacking in most applications, with few exceptions.

## Acknowledgements

The article was supported by the Spanish Ministry of Science, Innovation and Universities/FEDER (grant RTI2018-095518-B-C21), the Spanish Ministry of Health (grant CIBER CB06/02/1002) and the Catalan Government (grant 2017SGR656). Acknowledgements are due to Josep Antoni Martín Fernández, Michael Greenacre and Juan José Egozcue for their comments on previous drafts, to Karel Hron, Michael Greenacre, John Bacon-Shone, and Raimon Tolosana-Delgado for fruitful discussions at the 8th International Workshop on Compositional Data Analysis, and to Berta Ferrer-Rosell and Esther Martínez Garcia, for keeping asking “is that effect large or small?” in all articles we co-authored.

## References

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

Aitchison, J. (1983). Principal component analysis of compositional data. *Biometrika*, 70, 57–65.

Aitchison, J. (1986). *The Statistical Analysis of Compositional Data. Monographs on Statistics and Applied Probability*. Chapman and Hall, London.

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

Barceló-Vidal, C. and Martín-Fernández, J.A. (2016). The mathematics of compositional analysis. *Austrian Journal of Statistics*, 45, 57–71.

Boogaart, K.G. Van den and Tolosana-Delgado, R. (2013). *Analyzing Compositional Data with R*. Springer, Berlin.

Buccianti, A., Mateu-Figueras, G. and Pawlowsky-Glahn, V. (2006). *Compositional Data Analysis in the Geosciences: from Theory to Practice*. Geological Society, London.

Coenders, G. (2019). Compositional explanatory variables: which are the differences between alr and pivot coordinates? *8th International Workshop on Compositional Data Analysis CoDaWork 2019*. Terrassa, Spain, 3-8 June 2019. https://doi.org/10.13140/RG.2.2.22987.44325

Coenders, G. and Ferrer-Rosell, B. (2020). Compositional data analysis in tourism. Review and future directions. *Tourism Analysis*, 25, 153–168.

Coenders, G., Martín-Fernández, J.A. and Ferrer-Rosell, B. (2017). When relative and absolute information matter. Compositional predictor with a total in generalized linear models. *Statistical Modelling*, 17, 494–512.

Combettes, P.L. and Müller, C.L. (2019). Regression models for compositional data: General log-contrast formulations, proximal optimization, and microbiome data applications. *arXiv*,1903.01050.

Dumuid, D., Pedišić, Ž., Stanford, T.E., Martín-Fernández, J.A., Hron, K., Maher, C.A., Lewis, L.K. and Olds, T. (2019). The compositional isotemporal substitution model: a method for estimating changes in a health outcome for reallocation of time between sleep, physical activity and sedentary behaviour. *Statistical Methods in Medical Research*, 28, 846-857.

Egozcue, J.J. and Pawlowsky-Glahn, V. (2005). Groups of parts and their balances in compositional data analysis. *Mathematical Geology*, 37, 795–828.

Egozcue J.J. and Pawlowsky-Glahn, V. (2019). Compositional data: the sample space and its structure. *TEST*, 28, 599–638.

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

Filzmoser, P., Hron, K. and Templ, M. (2018). *Applied Compositional Data Analysis with Worked Examples in R*. Springer, New York.

Fišerová, E. and Hron, K. (2011). On interpretation of orthonormal coordinates for compositional data. *Mathematical Geosciences*, 43, 455–468.

Greenacre, M. (2018). *Compositional Data Analysis in Practice*. Chapman and Hall/CRC press, New York.

Greenacre, M. (2019). Variable selection in compositional data analysis using pairwise logratios. *Mathematical Geosciences*, 51, 649–682.

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

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

Linares-Mustarós, S., Coenders, G. and Vives-Mestres, M. (2018). Financial performance and distress profiles. From classification according to financial ratios to compositional classification. *Advances in Accounting*, 40, 1–10.

Louzada, F., Shimizu, T.K.O. and Suzuki, A.K. (2019). The Spike-and-Slab Lasso regression modeling with compositional covariates: An application on Brazilian children malnutrition data. *Statistical Methods in Medical Research*. https://doi.org/10.1177/0962280219863817

Martín-Fernández, J.A., Pawlowsky-Glahn, V., Egozcue, J.J. and Tolosona-Delgado, R. (2018). Advances in principal balances for compositional data. *Mathematical Geosciences*, 50, 273–298.

Morais, J., Thomas-Agnan, C. and Simioni, M. (2018). Interpretation of explanatory variables impacts in compositional regression models. *Austrian Journal of Statistics*, 47, 1–25.

Müller, I., Hron, K., Fišerová, E., Šmahaj, J., Cakirpaloglu, P. and Vančáková, J. (2018). Interpretation of compositional regression with application to time budget analysis. *Austrian Journal of Statistics*, 47, 3–19.

Ortells, R., Egozcue, J.J., Ortego, M.I. and Garola, A. (2016). Relationship between popularity of key words in the Google browser and the evolution of worldwide financial indices. In: Martín-Fernández, J.A. and Thió-Henestrosa, S. (eds), *Compositional Data Analysis. Springer Proceedings in Mathematics & Statistics*, Vol. 187. Springer, Cham.

Palarea-Albaladejo, J. and Martín-Fernández, J.A. (2015). zCompositions—R package for multivariate imputation of left-censored data under a compositional approach. *Chemometrics and Intelligent Laboratory Systems*, 143, 85–96.

Pawlowsky-Glahn, V. and Buccianti, A. (2011). *Compositional Data Analysis. Theory and Applications*. Wiley, New York.

Pawlowsky-Glahn, V., Egozcue, J.J. and Tolosana-Delgado, R. (2015). *Modelling and Analysis of Compositional Data*. Wiley, Chichester.

Pindyck, R.S. and Rubinfeld, D.L. (1976). *Econometric Models and Economic Forecasts*. MacGraw-Hill, New York.

Quinn, T. and Erb, I. (2020). Interpretable log contrasts for the classification of health biomarkers: a new approach to balance selection. *mSystems*, 5, e00230–19.

Rivera-Pinto, J., Egozcue, J.J., Pawlowsky-Glahn, V., Paredes, R., Noguera-Julian, M. and Calle, L. (2018). Balances: a new perspective for microbiome analysis. *mSystems*, 3, e00053–18.

Solans, M., Coenders, G., Marcos-Gragera, R., Castelló, A., Gràcia-Lavedan, E., Benavente, Y., Moreno, V., Pérez-Gómez, B., Amiano, P., Fernández-Villa, T., Guevara, M., Gómez-Acebo, I., Fernández-Tardón, G., Vanaclocha-Espi, M., Chirlaque, M.D., Capelo, R., Barrios, R., Aragonés, N., Molinuevo, A., Vitelli-Storelli, F., Castilla, J., Dierssen-Sotos, T., Castaño-Vinyals, G., Kogevinas, M., Pollán, M. and Saez, M. (2019). Compositional analysis of dietary patterns. *Statistical Methods in Medical Research*, 28(9), 2834–2847.

Thió-Henestrosa, S. and Martín-Fernández, J.A. (2005). Dealing with compositional data: The freeware CoDaPack. *Mathematical Geology*, 37, 773–793.

Tolosana-Delgado, R. and Boogaart, K.G. Van den (2011). Linear models with compositions in R. In: Pawlowsky-Glahn, V. and Buccianti, A. (eds), *Compositional Data Analysis. Theory and Applications*. Wiley, New York.