WORKING PAPERS

N° 1057

December 2019

“Covariates impacts in compositional models and simplicial derivatives”

Joanna Morais, Christine Thomas-Agnan

Toulouse School of Economics

Covariates impacts in compositional models and simplicial derivatives

December 2, 2019

Joanna Morais [2]& Christine Thomas-Agnan[1]

[1] *Toulouse School of Economics, France*
[2] *Avisia, Bordeaux, France*

Abstract

In the framework of Compositional Data Analysis, vectors carrying relative information, also called compositional vectors, can appear in regression models either as dependent or as explanatory variables. In some situations, they can be on both sides of the regression equation. Measuring the marginal impacts of covariates in these types of models is not straightforward since a change in one component of a closed composition automatically affects the rest of the composition.

J. Morais, C. Thomas-Agnan and M. Simioni [*Austrian Journal of Statistics,* 47(5), 1-25, 2018] have shown how to measure, compute and interpret these marginal impacts in the case of linear regression models with compositions on both sides of the equation. The resulting natural interpretation is in terms of an elasticity, a quantity commonly used in econometrics and marketing applications. They also demonstrate the link between these elasticities and simplicial derivatives.

The aim of this contribution is to extend these results to other situations, namely when the compositional vector is on a single side of the regression equation. In these cases, the marginal impact is related to a semi-elasticity and also linked to some simplicial derivative. Moreover we consider the possibility that a total variable is used as an explanatory variable, with several possible interpretations of this total and we derive the elasticity formulas in that case.

Key Words:compositional regression model, marginal effects, simplicial derivative, elasticity, semi-elasticity

1 Introduction and literature review

We consider regression models involving compositional vectors, i.e. vectors carrying relative information. When relative information is the focus, meaningful functions are functions of ratios of the vector's components therefore using traditional regression models in such cases is not correct. Regression models that respect the compositional nature of such data have been proposed in the literature, for example those introduced by Aitchison (1986) based on log-ratio transformations. Theory for inference in these models is developed for example in Pawlowsky-Glahn and Buccianti (2011), Van Den Boogaart and Tolosana-Delgado (2013), Pawlowsky-Glahn et al. (2015), and Filzmoser et al. (2018).

When the compositional vectors only appear as dependent variable, we will say that the model is of the 'Y-compositional' type (see e.g. Egozcue et al. (2012)). When they only appear as explanatory variables, we will say that the model is of the 'X-compositional' type (see e.g. Hron et al. (2012)). Finally, when they appear on both sides, we will say that the model is of the 'YX-compositional' type, see e.g. Kynclova et al. (2015), Chen et al. (2017), Morais et al. (2018a) and Morais et al. (2018b). A simplified version of the YX-compositional type is presented in Wang et al. (2013) and Morais et al. (2018b) later show that this model is equivalent to the so-called MCI (multiplicative competitive interaction) model introduced earlier in the marketing literature (Nakanishi and Cooper (1982)). It may also be relevant to include in the model the total of the different parts involved in the composition and we will consider each of the above models for the case with or without a total variable, see e.g. Coenders et al. (2017) and Coenders et al. (2015). Extensions with compositional functional predictors are presented in Sun et al. (2018), Bui et al. (2018) and Combettes and Muller (2019). Case studies using some of these models are presented in Hron et al. (2012), Trinh et al. (2018) for the X-compositional type, Morais et al. (2017) for the YX-compositional type'.

The focus of the present work is on the definition and interpretation of covariates impacts in these models, question addressed by much fewer papers. Muller et al. (2018) propose an interpretation for models of X-compositional or Y-compositional types based on using a specific type of orthogonal coordinates (called pivot coordinates, see e.g. Filzmoser et al. (2018)). Moreover they promote the replacement of the natural logarithm by the base-2 logarithm for enhancing the interpretability. The first drawback is that the resulting interpretation requires rerunning the model once for each component in the Y-compositional case. Moreover changes in log-ratios correspond to multiplicative increase (of the dependent or independent variables) in terms of relative dominance, i.e. the ratio of one component to the geometric mean of the others (while keaping all other log-ratios constant) which is not a very intuitive notion. This point of view is extended in Coenders and Pawlowsky-Glahn (2019) by considering changes in more general log-ratios leading to changes in any subset of components by a common factor (while reducing the remaining components accordingly).

Morais et al. (2018b) show that a natural interpretation tool in the YX-compositional model is the notion of elasticity. Indeed elasticities are commonly computed for the MCI model in the marketing literature (see Nakanishi and Cooper (1982)). Morais et al. (2018b) relate it to the notion of simplicial derivatives introduced in Egozcue et al. and Barcelo-Vidal et al. both in Pawlowsky-Glahn and Buccianti (2011).

With a different approach, Nguyen et al. (2018) bring a different light on the evaluation of these impacts by plotting the predicted components as a function of the explanatory but this graphing tool is limited to compositional dependent or explanatory variables with three components.

Finally, for the X-compositional model, Coenders and Pawlowsky-Glahn (2019) consider the introduction of the total variable among the explanatory and adapt the resulting interpretations, still in terms of log-ratio changes.

The objective of this paper is to extend Morais et al. (2018b) to the Y-compositional and the X-compositional models and to allow inclusion of the total variable in the models. In Section 2, we introduce notations and define the different specifications of the considered models. In Section 3, we demonstrate the equations linking elasticities or semi-elasticities (depending on considered model) with simplicial derivatives. Section 4 establishes the formulas for the elasticities and semi-elasticities in terms of model parameters in the simplex as well as in coordinate space. Finally, Sections 5 provides examples of applications to the X-compositional and to the Y-compositional models. We conclude in Section 6.

2 Compositional model specifications

Let us denote by $\check{\mathbf{X}} = (\check{X}_1, \cdots, \check{X}_{D_X})' \in \mathbb{R}_+^{D_X}$ a vector of D_X positive components corresponding to the components of a compositional vector expressed in original units: we call these components *volumes* as opposed to *shares*. For example, in the case studied in Morais et al. (2017), the volumes are numbers of cars sold during a given month by the different brands of cars whereas the shares represent the corresponding proportion of cars sold during that month by each brand relative to the other brands in the study. The closure of the vector $\check{\mathbf{X}}$ of volumes is the corresponding vector of shares

$$\mathbf{X} = \mathcal{C}(\check{X}_1, \cdots, \check{X}_{D_X})' = \left(\frac{\check{X}_1}{\sum_{i=1}^{D_X} \check{X}_i}, \cdots, \frac{\check{X}_{D_X}}{\sum_{i=1}^{D_X} \check{X}_i} \right)' = (X_1, \cdots, X_{D_X})'$$

and belongs to the simplex space $\mathcal{S}^{D_X}$ of positive vectors in $\mathbb{R}^{D_X}$ with sum equal to 1.

In some cases, it may be relevant to include in the regression model a variable measuring a total (hence not scale-invariant) which may be $\mathbf{T}(\mathbf{X})$ or $\mathbf{T}(\mathbf{Y})$. Pawlowsky-Glahn et al. (2015) argue that different formulas can be used for this total, for example one of the following two:

- Arithmetic total: $\mathbf{T}_A(\check{\mathbf{Z}}) = \sum_{i=1}^{D} \check{Z}_i$
- Geometric total: $\mathbf{T}_G(\check{\mathbf{Z}}) = (\prod_{i=1}^{D} \check{Z}_i)^{1/\sqrt{D}}$

The general principle of simplicial regression is to use transformations to transport the simplex space $\mathcal{S}^D$, equipped with the Aitchison geometry, into the Euclidian space $\mathbb{R}^{D-1}$ thus eliminating the simplex constraints problem. It is generally agreed upon to use log-ratio orthonormal coordinates (Pawlowsky-Glahn et al.

(2015)). We recall that to each $D \times (D-1)$ contrast matrix $\mathbf{V}$, constructed from an orthonormal basis of $\mathcal{S}^D$, corresponds an isometric transformation traditionally called $ilr_{\mathbf{V}}$. As advocated recently by Martín-Fernández (2019), we will rather use the name olr (orthogonal log ratio) for these transformations. We then have $\mathbf{z}^* = olr_{\mathbf{V}}(\mathbf{z}) = \mathbf{V}' \log(\mathbf{z})$, where the natural logarithm (denoted by log) is understood componentwise and the inverse transformation in $olr_{\mathbf{V}}^{-1}(\mathbf{z}^*) = \mathcal{C}(\exp(\mathbf{V}\mathbf{z}^*))$.

Using the traditional notations for the simplex operations (see Pawlowsky-Glahn et al. (2015)), the first row of Table 1 presents the formulation of the regression models explaining a collection of n i.i.d. random variables (simplex valued or not) by corresponding explanatory variables which may be simplex valued or not. The observations are indexed by t, $t = 1, \cdots n$. Because marginal effects only involve one explanatory at a time, if we had a model explaining a simplex valued variable by both types of explanatories, we would use the first and last columns of this table. The second row of Table 1 presents the corresponding model formulations in coordinate space for a given choice of olr transformation $olr_{\mathbf{V}}$. Parameters $\mathbf{a}^*$, $\mathbf{b}^*$ or $\mathbf{B}^*$ are then estimated by maximum likelihood in coordinate space where the regression is classical. Formulas to compute the corresponding parameter estimates in the simplex $\mathbf{a}$, $\mathbf{b}$ or $\mathbf{B}$ are available and it is known that these estimated parameters in the simplex are independent of the particular choice of $olr_{\mathbf{V}}$, i.e. of the particular choice of contrast matrix. In Table 1 the different formulations may involve a total variable $\mathbf{T}(\mathbf{X})$ or $\mathbf{T}(\mathbf{Y})$ and it is printed in grey to indicate that it is an option. Finally, we included in the formulations the particular case of the MCI model obtained when $D_X = D_Y$ and the matrix $\mathbf{B}^*$ is a multiple of the identity resulting in $\mathbf{B} \boxdot \mathbf{X} = b \odot \mathbf{X}$.

3 Semi-elasticities and simplicial partial derivatives

A marginal impact in a linear regression model is usually understood as the change in the expected value of the dependent variable $\mathbf{Y}$ induced by an additive increase of the explanatory of interest $\mathbf{X}$. In nonlinear models, it is rather understood as the infinitesimal equivalent, i.e. the derivative of the expected value of $\mathbf{Y}$ with respect to $\mathbf{X}$ and it may be non constant throughout the range of $\mathbf{X}$. In some nonlinear models, an elasticity or a semi-elasticity may be more natural. Indeed in a log-log model, if $\mathbb{E}(\log(Y))$ depends linearly on $\log(X)$, then the parameter of $\log(X)$ is equal to the above derivative and can be interpreted as the percent increase of $\mathbb{E}(Y)$ induced by a one percent increase of X. Finally, if the model is a semi-log model, the natural quantity is either the partial derivative of $\mathbb{E}(Y)$ with respect to $\log(X)$ (if the logarithm is on the right hand side of the regression equation) or symmetrically the partial derivative of $\mathbb{E}(\log(Y))$ with respect to X in the other case (if the logarithm is on the left hand side of the regression equation). This supports the idea that, in a simplicial regression model, one should turn attention to simplicial derivatives to

Table 1: Specifications of the compositional models and notations

	Y-compositional model	X-compositional model	YX-compositional model
in $\mathcal{S}^D$	$\mathbf{Y}_t = \mathbf{a} \oplus \check{X}_t \odot \mathbf{b} \oplus \epsilon_t$ $\oplus T(\check{\mathbf{Y}})_t \odot \mathbf{c}$	$\check{Y}_t = a + < \mathbf{b}, \mathbf{X}_t >_A + \epsilon_t$ $+cT(\check{\mathbf{X}})_t$	'CODA' model: $\mathbf{Y}_t = \mathbf{a} \oplus \mathbf{B} \boxdot \mathbf{X}_t \oplus \epsilon_t$ $\oplus T(\check{\mathbf{X}})_t \odot \mathbf{c}$ 'MCI' model: $\mathbf{Y}_t = \mathbf{a} \oplus b \odot \mathbf{X}_t \oplus \epsilon_t$ $\oplus T(\check{\mathbf{X}})_t \odot \mathbf{c}$
in $\mathbb{R}^{D-1}$	$\mathbf{Y}^*_t = \mathbf{a}^* + \mathbf{b}^* \check{X}_t + \epsilon^*_t$ $+\mathbf{c}^* T(\check{\mathbf{Y}})_t$	$\check{Y}_t =$ $a + \sum_{k=1}^{D_X-1} b^*_k X^*_{t,k} + \epsilon_t$ $+cT(\check{\mathbf{X}})_t$	'CODA' model: $\mathbf{Y}^*_t = \mathbf{a}^* + \mathbf{B}^* \mathbf{X}^*_t + \epsilon^*_t$ $+\mathbf{c}^* T(\check{\mathbf{X}})_t$ 'MCI' model: $\mathbf{Y}^*_t = \mathbf{a}^* + b\mathbf{X}^*_t + \epsilon^*_t$ $+\mathbf{c}^* T(\check{\mathbf{X}})_t$
Notations	$\mathbf{Y}_t, \mathbf{a}, \mathbf{b}, \epsilon_t \in \mathcal{S}^{D_Y}$, $\check{X}_t \in \mathbb{R}$ $\mathbf{Y}^*_t, \mathbf{a}^*, \mathbf{b}^*$, $\epsilon^*_t \in \mathbb{R}^{D_Y-1}$	$\mathbf{X}_t, \mathbf{b} \in \mathcal{S}^{D_X}$, $\check{X}_t, a, \epsilon_t \in \mathbb{R}$ $\mathbf{X}^*_t, \mathbf{b}^* \in \mathbb{R}^{D_X-1}$	$\mathbf{B} \in \mathbb{R}^{D_Y, D_X}$, $b \in \mathbb{R}$ $\mathbf{B}^* \in \mathbb{R}^{D_Y-1, D_X-1}$

evaluate the impacts of explanatory variables. Adapting the definition of derivative to the case where a function is simplex valued or is defined on the simplex stems from the fact that a change in a vector of shares cannot be just reduced to a change in one of the components since they are linked by their sum constraint: in other words, it is due to the fact that one of the variables lies in a subspace of $\mathbb{R}^D$.

More precisely, the quantities of interest are

- $\frac{\partial^{\oplus} \mathbb{E}^{\oplus} \mathbf{Y}}{\partial X}$ in the case of the Y-compositional model
- $\frac{\partial \mathbb{E} Y}{\partial^{\oplus} \mathbf{X}}$ in the case of the X-compositional model,
- $\frac{\partial^{\oplus} \mathbb{E}^{\oplus} \mathbf{Y}}{\partial^{\oplus} \mathbf{X}}$ in the case of the YX-compositional model,

where $\mathbb{E}^{\oplus}$ denotes the expectation of a simplex valued random variable (see Pawlowsky-Glahn and Buccianti (2011)) and where the symbol $\partial^{\oplus}$ indicates that the derivative should be understood in the simplicial derivative sense with respect to that variable (see Barcelo-Vidal et al. and Egozcue et al. in Pawlowsky-Glahn and Buccianti (2011)).

For the Y-compositional and X-compositional models, we are first going to express the relevant simplicial derivatives in terms of semi-elasticities.

Indeed, for the case of the X-compositional model, let us consider an homogeneous function of degree zero f defined from $\mathbb{R}_+^D$ to $\mathbb{R}$ inducing a function $\underline{f}$ on $\mathcal{S}^D$ by $\underline{f}(\mathbf{x}) = \underline{f}(\mathcal{C}(\check{\mathbf{x}})) = f(\check{x})$.

Propositions (13.10) and (13.13) in Barcelo-Vidal et al. in Pawlowsky-Glahn and Buccianti (2011), chapter 13, imply that the part-$\mathcal{C}$ derivatives of $\underline{f}$, which we denote here by $\frac{\partial \underline{f}(\mathbf{x})}{\partial^{\oplus}\mathbf{x}}$ are given by:

$$\frac{\partial \underline{f}(\mathbf{x})}{\partial^{\oplus}\mathbf{x}} = \frac{\partial f(\check{\mathbf{x}})}{\partial \log(\check{\mathbf{x}})} \tag{1}$$

Therefore the derivative of a function $\underline{f}$ of a simplex valued variable $\mathbf{x} = \mathcal{C}(\check{\mathbf{x}})$ corresponds to the ordinary semi-log derivative of the corresponding homogeneous function f of the volumes $\check{\mathbf{x}}$. Applying this result to the function expressing $\mathbb{E}\mathbf{Y}$ as a function of the share vector $\mathbf{X}$, we obtain the link between the simplicial derivative of this function and the semi-elasticity (or semi-log derivative) in the classical sense of the corresponding function of the volume vector $\check{\mathbf{X}}$.

Similarly, for the case of the Y-compositional model, for a simplex-valued function $\mathbf{f}$ of a real variable $x \in \mathbb{R}$, Theorem 12.2.6 in Egozcue et al. in Pawlowsky-Glahn and Buccianti (2011), chapter 12, implies that:

$$\frac{\partial^{\oplus}\mathbf{f}(x)}{\partial x} = \mathcal{C}\left(\exp\left(\frac{\partial \log \mathbf{f}(x)}{\partial x}\right)\right)',$$

where $\partial^{\oplus}\mathbf{f}$ denotes the simplicial derivative of $\mathbf{f}$ at x. This result links the simplicial derivatives of a simplex-valued function $\mathbf{f}$ to the semi-log derivatives (in the ordinary sense) of this function. Applying this result to the function expressing $\mathbb{E}^{\oplus}\mathbf{Y}$ as a function of X, we obtain the link between the simplicial derivative of this function and the semi-elasticity (or semi-log derivative) in the classical sense of $\mathbb{E}^{\oplus}\mathbf{Y}$ as a function of $\mathbf{X}$.

For the the YX-compositional model, Morais et al. (2018a) linked simplicial derivatives to elasticities in the case of a model without a total and in the particular case where the number of components D_Y of the Y composition is the same as that of the X composition (D_X). The limitation $D_Y = D_X$ in Morais et al. (2018a) was simply due to the particular application framework of this work but there is no additional mathematical difficulty to extend the result to $D_Y \neq D_X$. The corresponding formulas are recalled in Table 2 for completeness.

Finally, considering models including a total, one would need to define infinitesimal paths in the $\mathcal{T}$-space. Instead we consider three types of infinitesimal variations as described in Section 4.3.

For upcoming interpretations, it is interesting to consider first order Taylor approximations of such functions (of a simplex variable or simplex valued). For a function $\underline{f}$ from $\mathcal{S}^D$ to $\mathbb{R}$, consider as in Barcelo-Vidal et al. (in Pawlowsky-Glahn

Table 2: Simplicial derivative and (semi-)elasticities

Y-compositional model	X-compositional model	YX-compositional model
$\frac{\partial^{\oplus}\mathbb{E}^{\oplus}\mathbf{Y}}{\partial X}$ $= \mathcal{C}\left(\exp\left(\frac{\partial \log \mathbb{E}^{\oplus}\mathbf{Y}}{\partial X}\right)\right)'$	$\frac{\partial \mathbb{E}Y}{\partial^{\oplus}\mathbf{X}} = \frac{\partial \mathbb{E}Y}{\partial \log \check{\mathbf{X}}}$	$\frac{\partial^{\oplus}\mathbb{E}^{\oplus}\mathbf{Y}}{\partial^{\oplus}\mathbf{X}}$ $= \mathcal{C}\left(\exp\left(\frac{\partial \log \mathbb{E}^{\oplus}\mathbf{Y}}{\partial \log \check{\mathbf{X}}}\right)\right)'$

and Buccianti (2011)) the orthonormal basis $\mathbf{u_1}, \cdots, \mathbf{u_D}$ of $\mathcal{S}^D$ defined by

$$\begin{aligned} \mathbf{u_j} &= \left(\frac{D-1}{D}\right) \odot \mathcal{C}\mu_{\mathbf{j}} \\ \mu_{\mathbf{j}} &= \mathcal{C}(1, \cdots, 1, \exp(1), 1, \cdots, 1), \end{aligned}$$

where exp(1) is at the j^{th} position. From Barcelo-Vidal et al. (in Pawlowsky-Glahn and Buccianti (2011)), the first order Taylor's approximation is given by

$$\underline{f}(\mathbf{x} \oplus \delta \odot \mathbf{u_j}) \sim \underline{f}(\mathbf{x}) + \delta \frac{\partial f(\check{\mathbf{x}})}{\partial \log(\check{\mathbf{x}}_{\mathbf{j}})}. \tag{2}$$

This additive (in the simplex sense) increase of $\delta \odot \mathbf{u_j}$ corresponds to a multiplicative increase of the j^{th} component while holding constant all other ratios of remaining components. It is also equivalent in coordinate space, for a proper choice of olr transformation, to increase additively one olr component while keeping all others constant. To summarize, note that the increment is given by the product of δ by the classical semi-elasticity, i.e., a semi-log derivative in the ordinary sense of the corresponding function of the volumes. As we will see in Section 5, δ is proportional to the rate of change of x.

For a function $\mathbf{f}$ from $\mathbb{R}$ to $\mathcal{S}^D$, Egozcue et al. in Pawlowsky-Glahn and Buccianti (2011) obtain the following first order Taylor approximation for a small additive increase $\delta > 0$ of $x \in \mathbb{R}$

$$\mathbf{f}(x+\delta) \sim \mathbf{f}(x) \oplus \delta \odot \frac{\partial^{\oplus}\mathbf{f}(x)}{\partial x}$$

As in Morais (2017), let us go one step further in the approximation. Indeed,

$$\mathbf{f}(x) \oplus \delta \odot \frac{\partial^{\oplus}\mathbf{f}(x)}{\partial x} = \mathcal{C}(\mathbf{f}(x) \oplus \exp(\delta \frac{\partial \log \mathbf{f}(x)}{\partial x})).$$

Combining with a first order approximation of the exponential in a neighborhood of zero $\exp(\delta \frac{\partial \log \mathbf{f}(x)}{\partial x}) \sim 1 + \delta \frac{\partial \log \mathbf{f}(x)}{\partial x}$, we get the following approximation for the m^{th} component of $\mathbf{f}(x+\delta)$

$$\mathbf{f}_m(x+\delta) \sim \mathbf{f}_m(x)(1 + \delta \frac{\partial \log \mathbf{f}_m(x)}{\partial x}), \tag{3}$$

Taking the derivative of $\sum_{m=1}^{D} \mathbf{f}_m(x) = 1$, we get $\sum_{m=1}^{D} \mathbf{f}_m(x) e_m = 0$. Therefore the RHS vector in equation (3) belongs to $\mathcal{S}^D$. To summarize, note that in this case the percent increase of each component of $\mathbf{f}(x)$ is given by the classical semi-elasticity, i.e., a semi-log derivative in the ordinary sense of the function.

Finally for a function $\mathbf{f}$ from $\mathcal{S}_X^D$, to $\mathcal{S}_Y^D$, a similar approximation has been obtained in Morais (2017) for the particular case $D_X = D_Y$. Combining the above two results, we obtain easily that the Taylor approximation of a function $\mathbf{f}$ from $\mathcal{S}^{D_X}$ to $\mathcal{S}^{D_Y}$ is given by

$$\underline{f}_m(\mathbf{x} \oplus \delta \odot \mathbf{u_j}) \sim \underline{f}_m(x) \left(1 + \delta \frac{\partial \log \mathbf{f}_m(\check{x})}{\partial \log \check{x_j}}\right). \qquad (4)$$

showing that a percent increase of the components of $\mathbf{x}$, proportional to δ, induces a percent increase of each component of $\mathbf{f}(x)$ given by the classical elasticity of the corresponding component $\frac{\partial \log \mathbf{f}_m(\check{x})}{\partial \log \check{x_j}}$.

4 Elasticities and semi-elasticities in terms of model parameters

The aim is now to relate the elasticities/semi-elasticities of the previous section to the model parameters. The results of this section will be based on the following two lemmas which establish the formulas for the semi-log derivatives of an olr transformation and its inverse.

Lemma 4.1. *If $\mathbf{z}$ is a D-composition which is the closure of the vector $\check{\mathbf{z}}$ of $\mathbb{R}_+^D$, and if $\mathbf{z}^* = olr_V(\mathbf{z}) = \mathbf{V}' log(\mathbf{z})$ is the olr-transformed vector associated to the contrast matrix $\mathbf{V}$, then*

$$\frac{\partial olr_V(\mathbf{z})}{\partial \log \check{\mathbf{z}}} = \mathbf{V}'$$

This first lemma just results from the definition of the olr which is linear with respect to $\log \check{\mathbf{z}}$, and could be used for any other linear transformation.

Lemma 4.2. *If $\mathbf{z}$ is a D-composition which is the closure of the vector $\check{\mathbf{z}}$ of $\mathbb{R}_+^D$, and if $\mathbf{z}^* = olr_V(\mathbf{z}) = \mathbf{V}' log(\mathbf{z})$ is the olr-transformed vector associated to the contrast matrix $\mathbf{V}$, then*

$$\frac{\partial \log(olr_V{}^{-1}(\mathbf{z}^*))}{\partial \mathbf{z}^*} = \mathbf{W_z V},$$

where $\mathbf{W_z}$ is the $D \times D$ matrix with $(1 - z_i)$ for the i^{th} diagonal element and $-z_j$ elsewhere on the j^{th} column and where $z = olr_V^{-1}(\mathbf{z}^*)$.

To prove Lemma 4.2, using the formula for the inverse transformation of an olr, we see that one representent of the share vector $\mathbf{z}$ is given by $\check{\mathbf{z}} = \exp(olr_V{}^{-1}(\mathbf{z}^*))$. It is easy to see that $\log(\mathbf{z}) = \log(\check{\mathbf{z}}) - \log(S)\mathbf{n}$, where $S = T_A(\check{z}) = \sum_{i=1}^{D} \check{z}_i$ and $\mathbf{n}$ is the simplex unit vector. The derivative of the first term yields V because $\log(\check{\mathbf{z}}) = V\mathbf{z}^*$. S is linear in $\check{\mathbf{z}}$, and it is easy to see that its derivatives with respect to $\mathbf{z}^*$ are given by

$$\frac{\partial S}{\partial z_j^*} = \sum_{i=1}^{D} \frac{\partial \log(\check{z}_i)}{\partial z_j^*} \check{z}_i = \sum_{i=1}^{D} v_{ij} \check{z}_i.$$

Then we have

$$\frac{\partial \log(S)}{\partial z_j^*} = \frac{1}{S} \frac{\partial S}{\partial z_j^*} = \sum_{i=1}^{D} v_{ij} z_i$$

Combining first and second terms yields

$$\frac{\partial \log(z_i)}{\partial z_j^*} = v_{ij} - \sum_{i=1}^{D} v_{ij} z_i,$$

and it is then enough to check that this is the general term of the matrix $\mathbf{W_z V}$. If we define $\mathbf{W_z^*} = \mathbf{W_z V}$, note that $\mathbf{W_z^* V'} = \mathbf{W_z}$ (will be used later on).

4.1 Semi-elasticities for Y-compositional models and X-compositional models

In the case of Y-compositional and X-compositional models, the natural tool is semi-elasticities. However the formulas differ in the two cases:

- X-compositional case: $SE(Y, \check{\mathbf{X}}) = \frac{\partial \mathbb{E} Y}{\partial \log \check{\mathbf{X}}}$
- Y-compositional case: $SE(\mathbf{Y}, \check{X}) = \frac{\partial \log \mathbb{E}^{\oplus} \mathbf{Y}}{\partial X}$

Let us denote by V^X, respectively V^Y, the contrast matrices used for X, respectively Y. The computation in the X-compositional case uses Lemma 4.1. Indeed, for $j = 1, \cdots, D_X$

$$\frac{\partial \mathbb{E} Y}{\partial \log \check{X}_j} = \sum_{k=1}^{D_X - 1} \frac{\partial \mathbb{E} Y}{\partial X_k^*} \frac{\partial X_k^*}{\partial \log \check{X}_j} = \sum_{k=1}^{D_X - 1} b_k^* \mathbf{V_{jk}^X} \tag{5}$$

This result is presented in matrix form in Table 3. Note that this semi-elasticity is constant throughout observations.

The computation in the Y-compositional case uses Lemma 4.2 since $\mathbb{E}^{\oplus}\mathbf{Y} = olr_V^{-1}(\mathbb{E}\mathbf{Y}^*)$. We have

$$\frac{\partial \log \mathbb{E}^{\oplus} \mathbf{Y}}{\partial \check{X}} = \frac{\partial \log \mathbb{E}^{\oplus} \mathbf{Y}}{\partial \mathbb{E} \mathbf{Y}^*} \frac{\partial \mathbb{E} \mathbf{Y}^*}{\partial \check{X}} = \mathbf{W_z^* b^*} = \mathbf{W_z^* V^{Y'}} \log \mathbf{b} = \mathbf{W_z} \log \mathbf{b},$$

where $\mathbf{z} = olr_V{}^{-1}(\mathbb{E}(olr_V \mathbf{Y})) = \mathbb{E}^{\oplus}\mathbf{Y}$. Note that in this case the result depends upon the observation t through the dependence of $\mathbf{W_z}$ on $\mathbf{z}_t = \mathbb{E}^{\oplus}\mathbf{Y}_t$.

4.2 Elasticities for the YX-compositional model

For the YX-compositional model, Morais et al. (2018a) have obtained the expressions of the elasticities when the dimension of the Y composition is the same as that of the X composition. Let us extend this result to the case $D_X \neq D_Y$ using the above two lemmas.

We can see the relationship between $\log \check{\mathbf{X}}$ and $\log \mathbb{E}^{\oplus}\mathbf{Y}$ as the composition of three functions (listed from inside to outside)

- the function which maps $\log \check{\mathbf{X}} \in \mathbb{R}^{+D_X}$ to $\mathbf{X}^* \in \mathcal{S}^{D_X}$
- the function which maps $\mathbf{X}^* \in \mathcal{S}^{D_X}$ to $\mathbb{E}\mathbf{Y}^* \in \mathbb{R}^{+D_Y}$
- the function which maps $\mathbb{E}\mathbf{Y}^* \in \mathbb{R}^{+D_Y}$ to $\log \mathbb{E}^{\oplus}\mathbf{Y} \in \mathcal{S}^{D_Y}$

Using the generalized chain rule for functions of several variables which states that the Jacobian matrix of the composite function is the product of the Jacobian matrices of the composed functions evaluated at appropriate points, we get

$$\frac{\partial \log \mathbb{E}^{\oplus}\mathbf{Y}}{\partial \log \check{\mathbf{X}}} = \frac{\partial \log \mathbb{E}^{\oplus}\mathbf{Y}}{\partial \mathbb{E}\mathbf{Y}^*} \frac{\partial \mathbb{E}\mathbf{Y}^*}{\partial \mathbf{X}^*} \frac{\partial \mathbf{X}^*}{\partial \log \check{\mathbf{X}}} \tag{6}$$

The rightmost term on the right hand side of (6) is obtained using Lemma 4.1:

$$\frac{\partial \mathbf{X}^*}{\partial \log \check{\mathbf{X}}} = \mathbf{V}^{\mathbf{X}\prime}.$$

The central term yields the matrix $\mathbf{B}^*$ of parameters in coordinate space since the relationship between $\mathbb{E}\mathbf{Y}^*$ and $\mathbf{X}^*$ is linear. The leftmost term on the right hand side is obtained using Lemma 4.2:

$$\frac{\partial \log \mathbb{E}^{\oplus}\mathbf{Y}}{\partial \mathbb{E}\mathbf{Y}^*} = \mathbf{W}^*_{\mathbf{z}}\mathbf{V}^{\mathbf{Y}},$$

where $\mathbf{z} = \mathbb{E}^{\oplus}\mathbf{Y}$. We finally get

$$\frac{\partial \log \mathbb{E}^{\oplus}\mathbf{Y}}{\partial \check{\mathbf{X}}} = \mathbf{W}^*_{\mathbf{z}}\mathbf{V}^{\mathbf{Y}}\mathbf{B}^*\mathbf{V}^{\mathbf{X}\prime} = \mathbf{W}^*_{\mathbf{z}}\mathbf{B}. \tag{7}$$

Note that the result is again observation dependent. Table 3 summarizes the different formulas for semi-elasticities and elasticities for the three types of models as a function of parameters estimates, in the simplex or in coordinate space.

4.3 Models including a total

The presence of the total variable has to be taken into account in the partial impact measure computations. We consider including among the explanatory variables

- a total of $\mathbf{Y}$ in the Y-compositional model (model A)
- a total of $\mathbf{X}$ in the X-compositional model (model B)
- a total of $\mathbf{X}$ and/or a total of $\mathbf{Y}$ in the YX-compositional model (model C)

Table 3: (Semi-)elasticities without total

Y-compositional model	X-compositional model	YX-compositional model
$\frac{\partial \log \mathbb{E}^{\oplus}\mathbf{Y}}{\partial X} = \mathbf{W}^{*}_{\mathbb{E}\oplus\mathbf{Y}}\mathbf{b}^{*}$ $= \mathbf{W}_{\mathbb{E}\oplus\mathbf{Y}} \log \mathbf{b}$	$\frac{\partial \mathbb{E}(\check{Y})}{\partial \log \check{\mathbf{X}}} = \mathbf{V}^{\mathbf{X}}\mathbf{b}^{*}$ $= \mathbf{V}^{\mathbf{X}}\mathbf{V}^{\mathbf{X}'} \log \mathbf{b}$	'CODA' Model $\frac{\partial \log \mathbb{E}^{\oplus}\mathbf{Y}}{\partial \log \check{\mathbf{X}}} = \mathbf{W}^{*}_{\mathbb{E}\oplus\mathbf{Y}}\mathbf{B}^{*}\mathbf{V}^{\mathbf{X}'} = \mathbf{W}_{\mathbb{E}\oplus\mathbf{Y}}\mathbf{B}$ 'MCI' Model $\frac{\partial \log \mathbb{E}^{\oplus}\mathbf{Y}}{\partial \log \check{\mathbf{X}}} = \mathbf{W}_{\mathbb{E}\oplus\mathbf{Y}} b$

Notations $\mathbf{W}_{(D_Y,D_Y)}$ is a matrix with $(1 - Y_i)$ on the diagonal and $-Y_i$ elsewhere on the ith row.

The right hand side of model equations from Table 1 are modified as follows

- model A: add $\oplus T(\mathbf{Y_t}) \odot c$, where c is the parameter corresponding to the total effect
- model B: add $+dT(\mathbf{X}_t)$, where c is the parameter corresponding to the total effect
- model C: add $\oplus T(\mathbf{Y_t}) \odot c \oplus T(\mathbf{X_t}) \odot d$, where c and d are the parameters corresponding to the two total effects.

In the presence of a total, as mentioned in Section 3, we need to distinguish three types of infinitesimal variations for a compositional variable, let us call it $\mathbf{Z}$ because it will be $\mathbf{X}$ or $\mathbf{Y}$ as the case may be. The three types are as follows

- Type 1: the total $T(\mathbf{Z})$ remains constant and we look at infinitesimal variations of the composition $\mathbf{Z}$. Such variations correspond to considering derivatives in the direction of one of the unitary vectors of an orthonormal basis of $\mathcal{S}^{D_Z}$. With a proper choice of basis and of contrast matrix as in Hron et al. (2012), this corresponds to an infinitesimal change in one component, along a linear path in the simplex, keeping all but the first ILR constant.
- Type 2: the composition $\mathbf{Z}$ remains constant while the total is subject to an infinitesimal variation. Such variations correspond to considering ordinary derivatives with respect to the total $T(\mathbf{Z})$.
- Type 3: one of the components of $\mathbf{Z}$ varies together with the total $T(\mathbf{Z})$.

Type 1 variations A type 1 variation of $\mathbf{Y}$ would have no meaning in model A and in model C. In model B, the impact of a type 1 variation of $\mathbf{X}$ with fixed total can be computed as in the X-compositional model in Table 3. The impact of a type 1 variation of $\mathbf{X}$ in model C can be computed as in the YX-compositional model in Table 3.

Type 2 variations Type 2 variations correspond to ordinary derivatives with respect to the total. For models A and C, a type 2 variation for $\mathbf{Y}$ would have no meaning.

For model B, a type 2 variation of $\mathbf{X}$ results in an ordinary derivative

$$\frac{\partial \mathbb{E} Y}{\partial T} = c \tag{8}$$

In model C, a type 2 variation of $\mathbf{X}$ can be computed as in a Y-compositional model treating the total $T(\mathbf{X})$ as an ordinary variable and computing the derivative with respect to the total.

Type 3 variations First of all, type 3 variations only make sense for $\mathbf{X}$. Therefore we can't consider type 3 variations in model A.

Moreover, evaluating the effect of the variation of $\mathbf{X}$ or of $T(\mathbf{X})$ is equivalent since they are linked together, therefore one of the two formulas is enough.

For type 3 variations of $\mathbf{X}$, since both total and composition vary, the easiest way out is to express the dependent as a function of the volumes and use ordinary derivatives of the ensuing function of the volumes.

In model B, for computing the effect of a type 3 variation of $\mathbf{X}$, we need to adapt equation (5) adding an extra term taking into account the fact that the total depends upon the volumes and we get

$$\frac{\partial \mathbb{E} Y}{\partial \log \check{X}} = \mathbf{V}^{\mathbf{X}} \mathbf{b}^{*} + \frac{\partial \log \mathbb{E} \mathbf{Y}}{\partial T} \frac{\partial T}{\partial \log \check{X}} = \mathbf{V}^{\mathbf{X}} \mathbf{b}^{*} + d \frac{\partial T}{\partial \log \check{X}} \tag{9}$$

This result shows that the derivatives of the total with respect to the volumes play a role in the final expression of this semi-elasticity (hence we get a different formula for an arithmetic or a geometric total).

In model C, for a type 3 variation of $\mathbf{X}$, the derivative with respect to $\mathbf{X}$ of the first term $\mathbf{B} \boxdot \mathbf{X}$ is obtained as in the YX-compositional model without total and and the derivative of the second term $T(\mathbf{X}) \odot d$ is obtained as in the X-compositional model with a $T(\mathbf{X})$ total (equation (7)) yielding overall

$$\frac{\partial \log \mathbb{E}^{\oplus} Y}{\partial \log \check{X}} = \mathbf{W}_{\mathbb{E}^{\oplus} \mathbf{Y}} (\mathbf{B} + \log(d) \frac{\partial T}{\partial \log \check{X}}) \tag{10}$$

Once again, the result involves the the derivatives of the total with respect to the volumes.

5 Illustration

Let us give two toy examples of interpretation to illustrate our approach. We focus on the X-compositional and the Y-compositional models since the case of the YX-compositional model was already illustrated in Morais et al. (2018a).

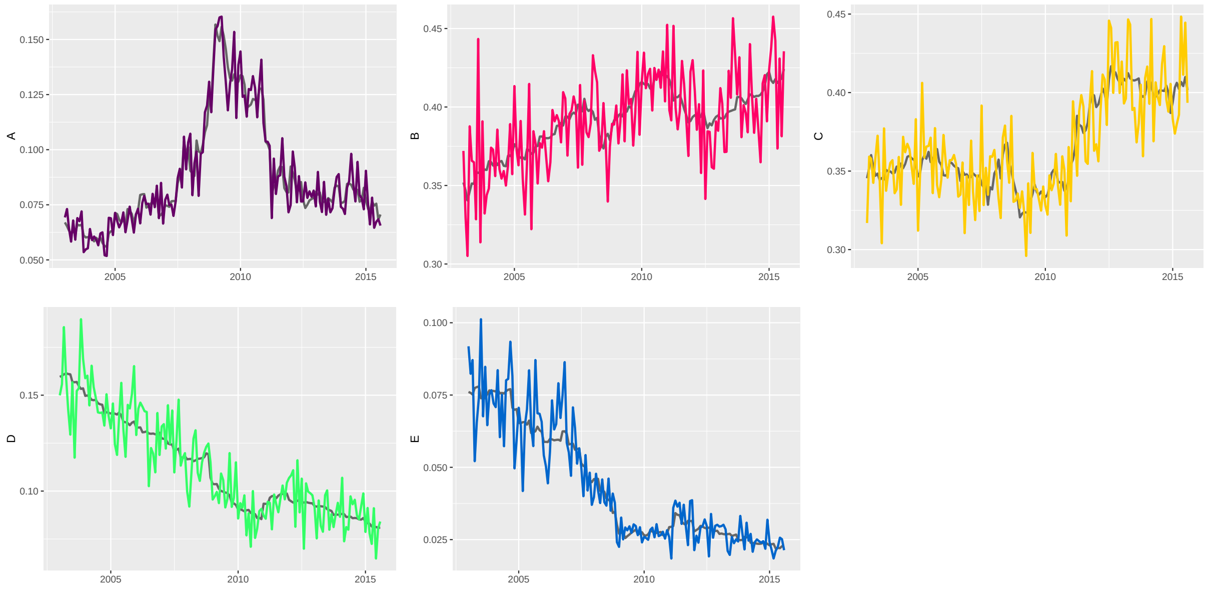

Figure 1: Observed (in color) and predicted (in grey) segments shares along time

5.1 Economic context and automobile market segment shares: Y-compositional model

In Morais J. (2020), the relationship between the socioeconomic context on the demand of new cars by segments is investigated with a data set coming from the French Renault company for market shares and from publicly available data bases. The data coming from Renault has been blurred with a small noise for confidentiality reasons. The automobile market is divided into five segments, from the smallest vehicles (A segment) to the largest vehicles (E segment). The available explanatory variables are consumption expenditure, an economic sentiment indicator, Gross Fixed Capital Formation of household, Gross Domestic Product, diesel price and short term interest rate. The data is recorded monthly from 2003 to 2015 (167 observations). The model explaining the market shares of each segment by the above explanatory is therefore a Y-compositional model in our terminology. We use the following sequential binary partition: B versus A, C versus A and B, D versus A, B and C, and E versus A, B, C and D to construct an orthonormal basis of the simplex and an associated olr transformation. Figure 5.1 displays the observed and predicted segments shares along time and we can see that the compositional model catches the general tendency, but not all the variance of this data. Table 5.1 contains the average semi-elasticities of segments shares with respect to GDP.

Let us interpret for example the effect of a small increase of GDP on the small cars (A segment) market shares. From formula (3), a small additive increase $\delta = 1$ billion euros (this amount representing 0.6% of the average monthly GDP) results on average in a multiplicative increase of 0.0028 % of the A segment market share.

Table 4: Average semi-elasticities of segments shares with respect to GDP

	$\overline{SE(\mathbf{S}_t, GDP_t)}$
A	2.88e-05
B	-0.17e-05
C	-0.96e-05
D	0.99e-05
E	1.18e-05

Instead of focusing on average elasticities, we could concentrate on a given point in time and compute the impact on the whole share vector of such a small increase in GDP. We could then check easily that the new shares vector is indeed in the simplex.

5.2 French GDP and job market: X-compositional model

In this second illustration, we are interested in the impact on French GDP of the structure (composition) and the volume (total) of the French job market in the three main sectors of activity: Agriculture (primary), Industry (secondary), and Services (tertiary). GDP is expressed in million euros (current price) and total employment in thousands of people. The data is collected quarterly from 2004 to 2018 [1] We use the olr transformation corresponding to the sequential binary partition: Agriculture versus Industry and Services, and Industry versus Services. We consider the model explaining the GDP as a function of total employment and the two olr coordinates associated to the above olr transformation. It is therefore an X-compositional model including a total, in this case the simple arithmetic total employment. Table 5.2 reports the semi-elasticities of GDP with respect to the three sectors at the mean value of composition corresponding to 788, 9196 and 19385 thousand employees for respectively Agriculture, Industry and Services. To apply formula (2), we consider a small $\delta > 0$ and a variation of $\oplus\delta \odot \mathbf{u_j}$ of $\mathbf{x}$, where $\mathbf{u_j}$ is the unit vector in the direction of the component Services. This variation of $\mathbf{x}$ is equivalent, when δ is small, to a relative variation of $\sqrt{3/2}\delta$ (i.e. multiplying $\mathbf{x}$ by $1 + \sqrt{3/2}\delta$.) The factor $\sqrt{3/2}$ is $\sqrt{D_X/D_X - 1}$ in the general case, corresponding to $\log(u_j)$ in the Taylor expansion in Barcelo-Vidal et al. in Pawlowsky-Glahn and Buccianti (2011) . Taking $\delta = 0.01\%$ results in an increase of around $\sqrt{3/2} * 19385 * 0.0001 = 2450$ people of the Services employment while the ratio between Agriculture and Industry employments remain constant, and the model predicts that the GDP should increase by 84 million euros. The marginal effect of the size of the job market, assuming that its composition stays the same is obtained by the parameter estimate of total employment in the model, which is equal to 26.52. When total employment increases by 1000 people, the GDP tends to increase by 26.5 millions. Note that using a base 2 logarithm as in Muller et al. (2015) is not usefull in our approach and would rather introduce an unnecessary constant.

[1] https://data.oecd.org/emp/employment-by-activity.htm

Table 5: Semi-elasticities of GDP with respect to employment sectors composition

	$SE(GDP, \check{\overline{\mathbf{EmplSect}}})$
AGR	-10157.26
INDU	-51706.00
SERV	841030.75

6 Conclusion

This contribution highlights the fact that elasticities or semi-elasticities are well-adapted to interpret the impacts of explanatories in all types of compositional regression models. It also links these elasticities or semi-elasticities to the simplicial derivatives of the expected response with respect to the considered explanatory variable. The models may contain compositional variables on the right hand side and/or on the left hand side of the regression equation, and may contain or not total variables (relative to the dependent or the explanatory variables). Further work should be done about confidence intervals for (semi-)elasticities which can be computed by the Delta method, or simply using a bootstrap approach.

An alternative but more complex tool used in Wang et al. (2013) and in Morais et al. (2018a) is the elasticity of a ratio of shares. In the framework of an MCI model, it would directly correspond to a parameter of the model, which is attractive, but relates to a change rate of a ratio of components and not of a single component and therefore is more difficult to vulgarize.

Acknowledgements

We acknowledge funding from ANR under grant ANR-17-EUR-0010.

References

Aitchison, J. (1986). The statistical analysis of compositional data. Monographs on statistics and applied probability. Chapman and Hall.

Bui, T. T. T., J. Loubes, L. Risser, and P. Balaresque (2018). Distribution regression model with a reproducing kernel hilbert space approach. arXiv preprint arXiv:1806.10493.

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

Coenders, G., B. Ferrer-Rosell, G. Mateau-Figueras, and V. Pawlowsky-Glahn (2015). Manova of compositional data with a total. CODAWORK2015.

Coenders, G., J. A. Martín-Fernández, and B. Ferrer-Rosell (2017). When relative and absolute information matter: compositional predictor with a total in generalized linear models. Statistical Modelling 17(6), 494–512.

Coenders, G. and V. Pawlowsky-Glahn (2019). On interpretations of tests and effect sizes in regression models with a compositional predictor.

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

Egozcue, J. J., J. Daunis-I-Estadella, V. Pawlowsky-Glahn, K. Hron, and P. Filzmoser (2012). Simplicial regression. the normal model. Journal of applied probability and statistics.

Filzmoser, P., K. Hron, and M. Templ (2018). Applied compositional data analysis. With Worked.

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

Kynclova, P., P. Filzmoser, and K. Hron (2015). Modeling compositional time series with vector autoregressive models. Journal of Forecasting 34(4), 303–314.

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

Morais, J. (2017). Impact of media investments on brands' market shares: a compositional data analysis approach. Ph. D. thesis, Toulouse School of Economics (TSE).

Morais, J., C. Thomas-Agnan, and M. Simioni (2017). Impact of advertising on brand's market-shares in the automobile market: a multi-channel attraction model with competition and carryover effects.

Morais, J., C. Thomas-Agnan, and M. Simioni (2018a). Interpretation of explanatory variables impacts in compositional regression models. Austrian Journal of Statistics 47(5), 1–25.

Morais, J., C. Thomas-Agnan, and M. Simioni (2018b). Using compositional and dirichlet models for market share regression. Journal of Applied Statistics 45(9), 1670–1689.

Morais J., Thomas-Agnan, C. (2020). Impact of the economic context on the automobile market segment shares: a compositional approach. Working Paper.

Muller, I., K. Hron, E. Fiserova, J. Smahaj, P. Cakirpaloglu, and J. Vancakova (2015). Time budget analysis using logratio methods.

Muller, I., K. Hron, E. Fiserova, J. Smahaj, P. Cakirpaloglu, and J. Vancakova (2018, Feb.). Interpretation of compositional regression with application to time budget analysis. Austrian Journal of Statistics 47(2), 3–19.

Nakanishi, M. and L. G. Cooper (1982). Simplified estimation procedures for mci models. Marketing Science 1(3), pp. 314–322.

Nguyen, A., T. Laurent, C. Thomas-Agnan, and A. Ruiz-Gazen (2018). Analyzing the impacts of socio-economic factors on french departmental elections with coda methods.

Pawlowsky-Glahn, V. and A. Buccianti (2011). Compositional data analysis: Theory and applications. John Wiley & Sons.

Pawlowsky-Glahn, V., J. J. Egozcue, and D. Lovell (2015). Tools for compositional data with a total. Statistical Modelling 15(2), 175–190.

Pawlowsky-Glahn, V., J. J. Egozcue, and R. Tolosana-Delgado (2015). Modeling and Analysis of Compositional Data. John Wiley & Sons.

Sun, Z., W. Xu, X. Cong, and K. Chen (2018). Log-contrast regression with functional compositional predictors: Linking preterm infant's gut microbiome trajectories in early postnatal period to neurobehavioral outcome. arXiv preprint arXiv:1808.02403.

Trinh, H. T., J. Morais, C. Thomas-Agnan, and M. Simioni (2018). Relations between socio-economic factors and nutritional diet in vietnam from 2004 to 2014: New insights using compositional data analysis. Statistical methods in medical research, 0962280218770223.

Van Den Boogaart, K. G. and R. Tolosana-Delgado (2013). Analysing Compositional Data with R. Springer.

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