Pairwise share-ratio interpretations of compositional regression models

Lukas Dargel and Christine Thomas-Agnan

September 19, 2023

Abstract

The interpretation of regression models with compositional vectors as dependent and/or independent variables has been approached from different perspectives. The first approaches that appeared in the literature in the literature are done in coordinate space after some log-ratio transformation of the share vectors. Considering the fact that these models are non-linear with respect to classical operations of the real space, another approach has been proposed based on infinitesimal increments or derivatives understood in a simplex sense, leading to elasticities or semi-elasticities interpretations in the original space that have the advantage of being independent of any log-ratio transformations. After briefly reviewing these two points of view, we show that some functions of elasticities or semi-elasticities are constant throughout the sample observations, which makes them natural parameters for interpreting CoDa models. We derive approximations of share ratio variations and link them to these parameters leading to transformatio-free interpretations in the original shares space. We use a real data set on the French presidential election to illustrate each type of interpretation in detail.

1 Introduction

The interpretation of parameters in a regression model is an essential step to understand the impact of changes in an explanatory variable on the dependent variable. Let us first recall that, in a classical linear regression model explaining the dependent Y by a set of explanatory variables X, the conditional expected value $\mathbb{E}Y \mid X$ is a linear function of X. Therefore, for any specific explanatory variable X_k, we can understand the parameter β_{X_k} of X_k as the additive increment of $\mathbb{E}Y \mid X$ when X_k increases by 1 unit (finite increment), all other explanatory variables remaining fixed (ceteris paribus), or alternatively the derivative of $\mathbb{E}Y \mid X$ with respect to X_k (infinitesimal increment). In econometrics, the derivative interpretation is known as the marginal effect and the two points of view (finite and infinitesimal increments) coincide for linear models. However, CoDa regression models involve share vectors variables, also called compositional variables. For this reason, they are not linear for the classical vector space structures of the shares spaces, on the right hand side of the regression equation (when applicable) as well as on the left hand side (when applicable) and that is why the first interpretations in the literature are done in coordinate space after some log-ratio transformation of the share vectors. The absence of linearity on the left hand side (when the dependent is compositional) can be resolved by adapting the definition of expectation to compositional variables (see e.g. [Pawlowsky-Glahn et al., 2015]) as was already mentioned in [Morais and Thomas-Agnan, 2021]. As we will show, the absence of linearity on the right hand side (when the explanatory is compositional) can be resolved by considering linear increments of explanatory variables when linearity is understood with respect to the simplex geometry introduced by Aitchison and defined by the perturbation and powering operations.

In this work, our first aim is to discuss the different interpretations proposed so far for compositional models and show the advantage of the share space approach in contrast to the coordinate space approach. A second objective is to present a complete illustration of the elasticities/semi-elasticities interpretations of [Morais et al., 2018] on a real dataset about vote shares in an election. Using compositional data techniques for analyzing electoral data is natural and several references can be found as [Katz and King, 1999] and [Nguyen et al., 2022]. Finally, for the dependent composition case, our last purpose is to propose a new interpretation based on share ratios variations in the original share spaces.

Section 2 presents an electoral dataset that we use to illustrate our interpretations by modelling the election results and turnout rate of the 2022 French presidential election. After recalling the basics of CoDa regression, Section 3 reviews the classical interpretations of CoDa regression models in coordinate space and in the original space. Section 4 discusses approximations of changes of shares (additionally of share ratios) due to small linear variations of a share vector along a linear path in the simplex. This tool will be useful for manipulating linear variations of compositional explanatory variables as well as for interpreting the variations of dependent share vectors. Section 5 adapts conditional plots to CoDa regression. Section 6 argues that, for the scalar dependent case, all approaches agree but ours allows to consider more general increments of the explanatory compositional variables. Section 7 reviews and illustrates the infinitesimal approach of [Morais et al., 2018] and [Morais and Thomas-Agnan, 2021]. Finally Section 8 develops the interpretation approach based on pairwise share ratios in a model with a scalar dependent and some compositional covariates as well as in a model with a CoDa dependent.

2 Electoral dataset description

For our application, we will use a dataset on the French presidential election in 2022 and combine it with socio-economic data from the census. With an electoral sociology perspective, we can study the link between the shares of votes for each candidate/party in a given division of the territory and some corresponding socio-economic variables describing the territory subdivisions such as the shares of the different socio-professional categories, or the shares of different age groups in the population. The database of electoral results comes from the Ministry of Interior which is responsible for organizing the elections in France, and we concentrate on the first round. This database is openly available[1] and contains the following information for the 35245 French municipalities: the name and identifier of a given municipality, the number of registered voters, the number of actually recorded votes, the numbers of blank and invalid votes and the share of votes for each candidate. Table 1 presents for each of the nine candidates the total number of votes and the shares relative to the total of expressed votes. Besides the main candidates Macron, Mélenchon and Le Pen, we have decided to group the small candidates into two blocks: the “Left block” with candidates Jadot, Roussel, Hidalgo, Poutou and Arthaud and the “Right block” with candidates Zemmour, Pécresse, Lassalle and Dupont-Aignan.

Table 1: Results of the first round of the French 2022 presidential election

Candidate	Votes	Vote share
M. Emmanuel MACRON	9 783 058	27.85%
Ms. Marine LE PEN	8 133 828	23.15%
M. Jean-Luc MÉLENCHON	7 712 520	21.95%
M. Éric ZEMMOUR	2 485 226	7.07%
Ms. Valérie PÉCRESSE	1 679 001	4.78%
M. Yannick JADOT	1 627 853	4.63%
M. Jean LASSALLE	1 101 387	3.13%
M. Fabien ROUSSEL	802 422	2.28%
M. Nicolas DUPONT-AIGNAN	725 176	2.06%
Ms. Anne HIDALGO	616 478	1.75%
M. Philippe POUTOU	268 904	0.77%
Ms. Nathalie ARTHAUD	197 094	0.56%

[1] https://www.data.gouv.fr/fr/datasets/election-presidentielle-des-10-et-24-avril-2022-resultats-definitifs-du-1er-tour/

The socio-economic data we use comes from the national census that can be freely downloaded on the INSEE (French National Institute of Statistics and Economic Studies) website [2]. Since our models have an illustrative purpose, we focus on a single scalar explanatory variable: the population density and a single compositional explanatory variable: the share of employment in different professional categories (PC), which we will detail later.

Merging the two databases is not straightforward because the census tracts differ from the municipalities, which means that we first need to aggregate the census data. Additionally, due to the absence of data for certain overseas territories and for French citizens living abroad, we only run our analysis on the municipalities in metropolitan France (those in geographic Europe), leading to an overall 3.9% loss of expressed votes.

3 State of the art

The following three subsections provide a reminder of the current state of CoDa regression, a short review of the covariate impacts in classical regression models and in CoDa regression.

3.1 Reminder on CoDa regression

Compositional data are vectors, often called CoDa vectors, of non-negative components. When the objective of the research question is to concentrate on the role of the relative values of their components, they are often imposed a constant sum contraint of one in order to select a unique representer in the class of collinear vectors. The representers, therefore, belong to a simplex of $\mathbb{R}^D$ which is defined, for an integer D, by

$$\mathcal{S}^D = \left\{ \mathbf{u} = (u_1, \dots, u_D)' : u_m > 0, m = 1, ..., D; \sum_{m=1}^{D} u_m = 1 \right\}. \tag{1}$$

The log-ratio approach to the statistical analysis of CoDa vectors, initiated by [Aitchison, 1986], uses log-ratio transformations before applying classical statistical techniques to the transformed values. Let us recall the definition of the clr transformation, defined for a vector $\mathbf{u} \in \mathcal{S}^D$, by

$$\operatorname{clr}(\mathbf{u}) = \mathbf{G}_D \log \mathbf{u}, \tag{2}$$

where the logarithm of $\mathbf{u}$, understood componentwise, is a centered version of the vector of log-transformed components and where $\mathbf{G}_D = \mathbf{I}_D - \frac{1}{D}\mathbf{1}_D\mathbf{1}_D{}'$ with $\mathbf{1}_D$ being a D-vector of ones. The vector of shares (or parts, or components) are classically obtained by the closure of a vector $w \in \mathbb{R}^{D+}$ where the closure $\mathcal{C}$ is defined by

$$\mathcal{C}(\mathbf{w}) = \left(\frac{w_1}{\sum_{m=1}^{D} w_m}, \cdots, \frac{w_D}{\sum_{m=1}^{D} w_m} \right). \tag{3}$$

[Aitchison, 1986] equipped the simplex with a Hilbert space structure compatible with the objective of the CoDa analysis using the following operations:

1. the perturbation operation, denoted by $\oplus$, plays for $\mathcal{S}^D$ the role of the addition in $\mathbb{R}^D$:

$$\text{for} \quad \mathbf{u}, \mathbf{v} \in \mathcal{S}^D, \mathbf{u} \oplus \mathbf{v} = \mathcal{C}(u_1 v_1, \dots, u_D v_D),$$

2. the power operation, denoted by $\odot$, plays for $\mathcal{S}^D$ the role of the scalar multiplication in $\mathbb{R}^D$:

$$\text{for} \quad \lambda \in \mathbb{R}, \mathbf{u} \in \mathcal{S}^D \quad \lambda \odot \mathbf{u} = \mathcal{C}(u_1^\lambda, \dots, u_D^\lambda).$$

[2] https://www.insee.fr/fr/statistiques/6543200

The Aitchison inner product is given by

$$< \mathbf{u}, \mathbf{v} >_A = < \operatorname{clr}(\mathbf{u}), \operatorname{clr}(\mathbf{v}) >_E, \tag{4}$$

where the right hand side inner product is the standard inner product in $\mathbb{R}^D$. From the above definition, the clr transform is therefore an isometry between the simplex equipped with the Aitchison inner product and the classical Euclidean space $\mathbb{R}^D$. More complex transformations, (for example isometric log-ratio, denoted ilr) are often used but will not be needed in this work.

The product of a $D \times L$ matrix $\mathbf{B}$ by a vector $\boldsymbol{u} \in \mathcal{S}^L$ can be defined by $\mathbf{B} \boxdot \boldsymbol{u} = \operatorname{clr}^{-1}(\mathbf{B} \operatorname{clr}(\boldsymbol{u}))$ for any matrix $\mathbf{B}$ belonging to the set $\mathcal{A}_{DL}$ of $D \times L$ matrices with zero column sums and row sums (see generalized zero-sum property in [Ruiz-Gazen et al., 2023]). The definition of the clr transformation can also be extended to square matrices by $\operatorname{clr}(\mathbf{B}) = \mathbf{G}_D \mathbf{B} \mathbf{G}_D$ (see [Ruiz-Gazen et al., 2023]). For a simplex valued random variable $\mathbf{X}$, we use the following definitions for expectation and variance:

$$\begin{aligned} \mathbb{E}^{\oplus}\mathbf{X} &:= \operatorname{clr}^{-1}(\mathbb{E} \operatorname{clr}(\mathbf{X})) \qquad (5) \\ \mathbb{V}\mathrm{ar}^{\oplus}\mathbf{X} &:= \operatorname{clr}^{-1}(\mathbb{V}\mathrm{ar} \operatorname{clr}(\mathbf{X})). \qquad (6) \end{aligned}$$

The normal distribution on the simplex, which can be traced back to [Aitchison and Shen, 1980], is described in a more modern way in [Pawlowsky-Glahn et al., 2015]: $\mathcal{N}_{\mathcal{S}^D}(\boldsymbol{\mu}, \Omega)$ denotes the normal distribution of a simplex valued random variable $\mathbf{X}$ with $\boldsymbol{\mu} = \mathbb{E} \operatorname{clr}(\mathbf{X})$ and $\Omega = \mathbb{V}\mathrm{ar} \operatorname{clr}(\mathbf{X})$. The statistical analysis of CoDa vectors has first developped in the direction of descriptive and multivariate methods. Regression with CoDa vectors appeared in [Daunis-i Estadella et al., 2002]. The case of compositional explanatory variables is discussed for example in [Hron et al., 2012] and the case of compositional response for example in [Egozcue et al., 2012]. The case with CoDa vectors on both sides of the equation appears for example in [Chen et al., 2017]. Let us write the general equation for a CoDa regression model with compositional dependent variable $\mathbf{Y}$ and a mixture of classical and compositional explanatory variables using the simplex operations. The simplex equation of the model writes as follows

$$\mathbf{Y}_i = \boldsymbol{\alpha} \oplus (\bigoplus_{k=1}^{K_X} \mathbf{B}_k \boxdot \mathbf{X}_{ki}) \oplus (\bigoplus_{k=1}^{K_Z} Z_{ki} \odot \boldsymbol{\gamma}_k) \oplus \boldsymbol{\epsilon}_i \text{ with } \boldsymbol{\epsilon}_i \sim \mathcal{N}_{\mathcal{S}^{D_Y}}(\mathcal{C}(\mathbf{1}_{D_Y}), \mathbf{G}_{D_Y} \Sigma \mathbf{G}_{D_Y}), \tag{7}$$

where the parameter $\boldsymbol{\alpha}$ is a vector of $\mathcal{S}^{D_Y}$, the matrices of parameters $\mathbf{B}_k \in \mathcal{A}_{D_Y D_{X_k}}$, $k = 1, \ldots, K_X$, the $\mathbf{X}_{ki}$, $k = 1, \ldots, K_X$, are compositional explanatory variables in the simplex $\mathcal{S}^{D_{X_k}}$, the Z_{ki}, $k = 1, \ldots, K_Z$, are classical real-valued explanatory variables, and Σ is a positive definite symmetric $D_Y \times D_Y$ matrix.

For a real valued dependent variable, a regression model involving a combination of classical and compositional explanatory variables writes as follows:

$$Y_i = \alpha + \sum_{k=1}^{K_X} < \boldsymbol{\beta}_k, \mathbf{X}_{ki} >_A + \sum_{k=1}^{K_Z} \gamma_k Z_{ki} + \epsilon_i \text{ with } \epsilon_i \sim \mathcal{N}(0, \sigma^2), \tag{8}$$

with the same notations as above for the explanatory variables.

A more recent concept of compositional data analysis removing the constant sum constraint appears in see, e.g. [Ferrer-Rosell et al., 2016] and [Coenders et al., 2017] with the so-called $\mathcal{T}$-space introduced by [Pawlowsky-Glahn et al., 2015]. The absolute information involved in a so-called total variable $T(\mathbf{X})$ associated with any multivariate and positive explanatory variable $\mathbf{X}$ can be included as well on the right hand side of the regression equation among the classical variables.

The estimation of models (7) and (8) is obtained using multiple multivariate regression in a chosen ilr coordinate space and the result is independent of this choice. Formulas for the expression of the simplex parameters as a function of parameters in coordinate space are summarized in the paragraph "Transition formulas for log-ratio spaces" of the Annex.

Let us illustrate these regression models on the election data. A first model with a scalar dependent variable explains the turnout as a function of the logarithm of the population density (classical variable) and an ilr of the professional categories (CoDa variable). We want to point out that in this model the dependent variable is actually a rate, and could be treated as a 2 dimensional CoDa vector with a logit transformation. However, since our focus is just to illustrate the interpretations of all possible CoDa models we chose to continue with this imperfection instead of introducing a second data set. Moreover, in the range of most of the data, the relationship between the turnout rate and its logit is rather linear. A second model with compositional dependent variable explains the vote shares by the same explanatory variables.

The analysis of variance tables of the two models are respectively shown in Table 2 for the first one and Table 3 for the second. Note that while the constant term is not relevant in the anova for a univariate regression, it appears here because the dependent variable is multivariate and the "(Intercept)" refers to the vector $\boldsymbol{\alpha}$ in model (7). The two models display significant covariates.

Table 2: Analysis of variance table for scalar dependent model

	Df	Sum Sq	Mean Sq	F value	Pr(> F)
log(POP_DENSITY)	1	14328.87	14328.87	7799.13	0
ilr(PROFCAT)	3	43446.42	14482.14	7882.55	0
Residuals	34815	63963.50	1.84		

Table 3: Analysis of variance table for the compositional dependent model

	Df	Pillai	approx F	num Df	den Df	Pr(>F)
(Intercept)	1	0.847	38609	5	34811	0
log(POP_DENSITY)	1	0.195	1688	5	34811	0
ilr(PROFCAT)	3	0.184	455	15	104439	0
Residuals	34815					

3.2 Impact of a given covariate in a classical regression model

In this section, we focus on evaluating the impact of a given explanatory variable X on the dependent variable Y in a regression model explaining Y by X and by the rest of the explanatory variables denoted by $\mathbf{U}$. Loosely speaking, we want to measure the change of the expected value of Y due to an increment of X, while the other covariates remain unchanged (ceteris paribus). Depending on the nature of the model, different quantities will measure this impact. Two approaches are possible: considering a finite change or an infinitesimal change of X. In linear models, finite changes are adapted and the impact is then measured by the change of Y induced by a unit change of X. In nonlinear models, it is more adapted to replace finite changes by infinitesimal ones and to use the derivative of Y with respect to X_k, also called marginal effect.

When evaluating the impact of a particular X on Y, we will denote by

$$\tilde{Y}_i(x) := \mathbb{E}(Y_i \mid X = x, \mathbf{U}) \tag{9}$$

the value of the conditional expectation of Y at a given observation unit i given the value of all other explanatory variables, evaluated at the value $X = x$ of the considered variable X.

Let us first examine what happens for classical real variables Y and X in some simple models. In the classical linear regression model, we recall that the absolute change of Y induced by an increment of one unit of X ceteris paribus is simply equal to the parameter β_X of the variable X and also coincides with the derivative of $\tilde{Y}_i(x)$ with respect to x for any observation unit i. Let us now turn attention to models obtained by combining log transformations and a linear model. We first consider a model where $\mathbb{E}Y \mid X, \mathbf{U}$ is a linear function of the logarithm of X: $\mathbb{E}Y \mid X, \mathbf{U} = \alpha + \beta_X \log X + R$, where the term R contains the effect of other explanatory variables in $\mathbf{U}$. In this model it is clear that the parameter β_X can be viewed as the derivative $\frac{\partial \mathbb{E}\mathbf{Y}|X,\mathbf{U}}{\partial \log X}$ of $\mathbb{E}Y \mid X, \mathbf{U}$ with respect to $\log X$. In this case, for a small increment h of X, approximating $\exp(h) - 1$ by h, the increment of Y due to a relative increase of X of h units results in an additive increase of Y of $\beta_X h$. In econometrics, this parameter is called a semi-elasticity and it corresponds to a semi-logarithmic derivative. In this case, a marginal effect can be calculated by $\frac{\partial \mathbb{E}Y|X,\mathbf{U}}{\partial X} = \frac{\partial \mathbb{E}\mathbf{Y}|X,\mathbf{U}}{\partial \log X}\frac{\partial \log X}{\partial X}$ and is the ratio of the semi-elasticity to the value of X. In general, the marginal effects will be individual dependent, whereas the semi-elasticity is constant equal to β_X and therefore constitutes a natural parameter for this model.

Symmetrically, in a model where $\log \mathbb{E}Y \mid X, \mathbf{U}$ is a linear function of $\mathbf{X}$: $\log \mathbb{E}Y \mid X = \alpha + \beta_X X + R$, where the term R contains the effect of other explanatory variables in $\mathbf{U}$, the semi-logarithmic derivative $\frac{\partial \log \mathbb{E}Y|X,\mathbf{U}}{\partial X}$, also called semi-elasticity, is constant and equal to β_X. It allows approximating, for small increments of X, the relative increase of Y and the corresponding marginal effects, which are individual dependent. Once again, this semi-elasticity constitutes a natural parameter for this model.

In a log-log model such as $\log \mathbb{E}Y \mid X, \mathbf{U} = \alpha + \beta_X \log X + R$, where the term R contains the effect of other explanatory variables in $\mathbf{U}$, the logarithmic derivative $\frac{\partial \log \mathbb{E}\mathbf{Y}|X,\mathbf{U}}{\partial \log X}$ is constant and equal to β_X. It allows approximating for small relative increments of X the relative increase of Y.

Finally, let us recall what happens in a logistic model. The logit transformation ($\text{logit}(z) = \frac{z}{1-z}$) of $\mathbb{E}Y \mid X, \mathbf{U}$ being a linear function of the explanatory variables results in the fact that the natural parameter to interpret the logit model is given by the odds ratio which is constant and equal to the parameter whereas once again the corresponding marginal effect is observation dependent.

3.3 Impacts in CoDa regression models

Keeping these facts in mind, we now consider the case of CoDa regression models. The non linearity (in the classical sense) of these models and the compositional nature of some of their variables lead to additional difficulties. When the dependent is compositional, we need to adapt the notation in (9) because it then makes sense to use the compositional expectation of $\mathbf{Y}$:

$$\tilde{\mathbf{Y}}_i(x) := \mathbb{E}^{\oplus}(\mathbf{Y}_i \mid X = x, \mathbf{U}) \tag{10}$$

Now, when the explanatory variable of interest $\mathbf{X}$ is compositional, the usual "all things equal" does not apply anymore since one cannot change a single component of $\mathbf{X}$. It is then natural for interpreting the impact of a compositional variable $\mathbf{X} \in \mathcal{S}^{D_X}$ to replace the classical increment by a linear increment in the simplex as will be defined in Section 4. [Morais and Thomas-Agnan, 2021] use the appropriate simplicial derivatives to compute an approximation of the corresponding impact due to a small simplex linear increment. After recalling some simple facts about linear increments in a simplex, we recall the first order Taylor approximations of functions whose arguments and/or outcome values belong to a simplex. In the framework of CoDa regression, we then recall how the simplicial derivatives expressions in the Taylor formulas are linked to the model parameters.

Impacts in CoDa regression models have been studied with different techniques and for all types of regression models. For the model with a scalar dependent variable and explanatory variables including compositional and classical ones, this issue is well presented in [Coenders and Pawlowsky-Glahn, 2020]. For the model with compositional dependent variables, the impact of a scalar explanatory is discussed in [Müller et al., 2016], [Trinh et al., 2018], [Nguyen et al., 2020]. For a model involving a compositional dependent and at least one compositional explanatory variable, an approach can be found in [Müller et al., 2016] and for a very basic CoDa regression model in [Wang et al., 2013]. Other discussions are in [Morais et al., 2018], [Morais and Thomas-Agnan, 2021] and [Thomas-Agnan et al., 2023]. The models with a scalar dependent variable are the easiest to treat and all discussions (even though more or less general) agree on the interpretations. However, for models involving compositional dependent variables, the question is more intricate and there are two different points of view. For example, [Müller et al., 2016] consider an interpretation based on finite increments of fixed size for the covariate and describe the results of the covariate change using pivot coordinates ending up in measuring the change in the dominance of one particular share of $\mathbf{Y}$ with respect to the geometric mean of the other ones. The change of $\mathbf{X}$ they consider is described in terms of ilr coordinates but does indeed correspond to one segment in a particular linear path as in Section 11 although it is not described in these terms. More recently [Morais et al., 2018] and later [Morais and Thomas-Agnan, 2021] introduce interpretations based on small increments of variable size. These result in interpretations allowing to measure the change in each share of $\mathbf{Y}$ due to a change of $\mathbf{X}$ along a linear path in the $\mathbf{X}$ simplex space.

4 Linear increments in the simplex

In order to describe the impact of changing a particular compositional explanatory variable, we first need to define the considered variation. Indeed for a compositional variable, it is not possible to envision changing one of its components while keeping the others fixed because their sum is constrained. Hence, we need to describe the path of $\mathbf{X}$ in its simplex and it seems natural to consider a linear path. It turns out that studying linear paths in the simplex will also be helpful to understand the changes in the dependent variable.

Given an initial point in a simplex $\boldsymbol{x}(0) \in \mathcal{S}^D$, a vector $\boldsymbol{u} \in \mathcal{S}^D$ describing a direction, and a real h not necessarily positive (a signed intensity), we define $\boldsymbol{x}(h) \in \mathcal{S}^D$ by

$$\boldsymbol{x}(h) = \boldsymbol{x}(0) \oplus h \odot \boldsymbol{u}. \tag{11}$$

When h ranges from $-\infty$ to $+\infty$, $\boldsymbol{x}(h)$ describes a simplex line. In practice, it can be interesting to normalize the direction vector: when $\boldsymbol{u}$ has unit length, the Aitchison distance between $\boldsymbol{x}(h)$ and $\boldsymbol{x}(0)$ is equal to the absolute value of h.

Let us first derive an approximation for small h of the relative share variations along the path (11). For the j^{th} share, the relative variation is given by α_j :

$$\alpha_j := \frac{x_j(h) - x_j(0)}{x_j(0)} \simeq h[\log(u_j) - \sum_{l=1}^{D} x_l(0)\log(u_l)] = h\sum_{l=1}^{D} x_l(0)\log(\frac{u_j}{u_l}) \tag{12}$$

A first direction vector of particular interest is the vector $\boldsymbol{e_m} = \mathcal{C}(1, \ldots, e, \ldots 1)$ where $e = \exp(1)$ is in the m^{th} position. This particular direction plays for the simplex $\mathcal{S}^D$ the same role as the m^{th} vector of the canonical basis of $\mathbb{R}^D$. A simplex line as (11) with direction $\boldsymbol{u} = \boldsymbol{e_m}$ connects the m^{th} vertex of the simplex with a point on the opposite side. The m^{th} vertex is approached when h tends to ∞ and the corresponding components approach zero everywhere except for component m which approaches one. The point on the opposite side of the simplex is obtained by replacing in $\boldsymbol{x}(0)$ the m^{th} component by 0 and it is approached when h tends to $-\infty$.

For the direction $\boldsymbol{u} = \boldsymbol{e_m}$, (12) yields the following approximations:

$$\text{if} \qquad l \neq m, \quad \alpha_l := \frac{x_l(h) - x_l(0)}{x_l(0)} \simeq -hx_m(0), \tag{13}$$

$$\text{if} \qquad l = m, \quad \alpha_m := \frac{x_m(h) - x_m(0)}{x_m(0)} \simeq h(1 - x_m(0)). \tag{14}$$

Turning now attention to the share ratio variations, it is easy to show that for all couples of components i and j ($\in [1, D]$) we have the following finite increment formula for a general direction $\mathbf{u}$:

$$\frac{x_j(h)}{x_l(h)} = \frac{x_j(0)}{x_l(0)} \exp(h \log \frac{u_j}{u_l}). \tag{15}$$

Equation (15) proves that the share ratio variations are totally driven by the corresponding share ratios $\frac{u_j}{u_l}$ of the direction vector. Moreover, we see that if two components j and l of the direction $\boldsymbol{u}$ are equal, then the ratio $\frac{x_j(h)}{x_l(h)}$ is not affected by the change. At the other extreme the share ratios of x will be more affected for couples of components u_j and u_l which are the most dissimilar.

If we now consider a very small value of h, we can make an approximation in (15) to obtain the infinitesimal share ratio increments

$$\frac{x_j(h)}{x_l(h)} \simeq \frac{x_j(0)}{x_l(0)}(1 + h \log \frac{u_j}{u_l}), \tag{16}$$

which can also be written in terms of relative share ratio variations as follows

$$\frac{\frac{x_j(h)}{x_l(h)} - \frac{x_j(0)}{x_l(0)}}{\frac{x_j(0)}{x_l(0)}} \simeq h \log \frac{u_j}{u_l}. \tag{17}$$

For the particular direction $\boldsymbol{e_m}$ we obtain the finite increments expression of the share ratios

$$\text{if} \qquad l \neq m, \quad \frac{x_m(h)}{x_l(h)} = \frac{x_m(0)}{x_l(0)} \exp(h) \tag{18}$$

$$\text{if} \qquad l, j \neq m, \quad \frac{x_j(h)}{x_l(h)} = \frac{x_j(0)}{x_l(0)}. \tag{19}$$

The corresponding infinitesimal share ratio increments are obtained by replacing $\exp(h)$ by $1+h$ in (19).

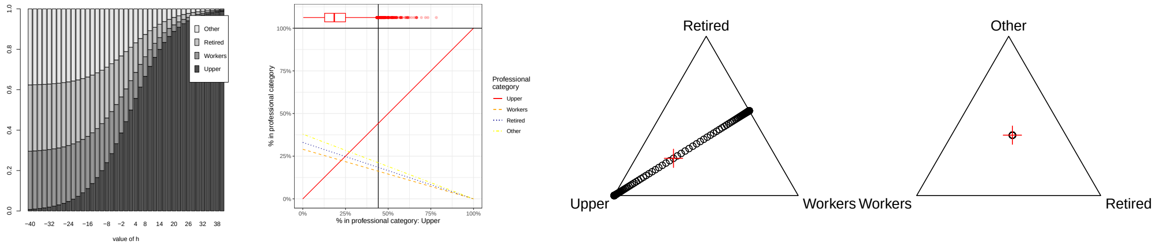

Figure 1: Change in the direction of the vertex "Upper"

Figure 1 illustrates in three different ways the change in the direction of the vertex "Upper" in the simplex of the four socio-professional categories. Moving in this direction means increasing the share of "Upper" while uniformly decreasing the other shares. The stacked barplot on the left represents, for h varying on a grid on the x-axis, the four corresponding shares along the linear path (see pages 28-29 of [Boogaart and Tolosana-Delgado, 2013]). The second plot is a simple scatterplot of the four categories shares as a function of the share of "Upper". The vertical line corresponds to the initial point $z(0)$ which is the municipality of Paris. The boxplot on top of this graph depicts the univariate distribution of the share of "Upper", showing that Paris has among the highest proportions of individuals in this professional category. Note that there are situations in which, on this second graph, some shares are multi-valued functions of the x-axis share. Both ternary diagrams are based on subcompositions with three out of four professional categories, and a cross indicates the initial point $z(0)$ in each of them. On the right ternary diagram, we can check that the subcomposition of the three vertices other than "Upper" is constant along the path, which corresponds to equation (19).

Similarly, Figure 2 illustrates a linear path in a general direction defined by $\mathbf{u} = (0.1, 0.5, 0.25, 0.15)$. Moving in this direction can be summarized by what happens for an increment of $h = 1$: multiplying the ratio of the second to the first component by $5 = \frac{u_2}{u_1}$, the ratio of the fourth to the third component by $0.6 = \frac{u_4}{u_3}$, etc. Since the highest ratio between the $\mathbf{u}$ components is the ratio of the first to the second component equal to 1/5, it is the share ratio of “Upper” to “Workers” which varies the most, which is reflected in the first three graphs of 2. Similarly, the ratio which fluctuates the least is the one of “Other” to “Retired” corresponding to a ratio of 5/3 of the corresponding components of $\mathbf{u}$.

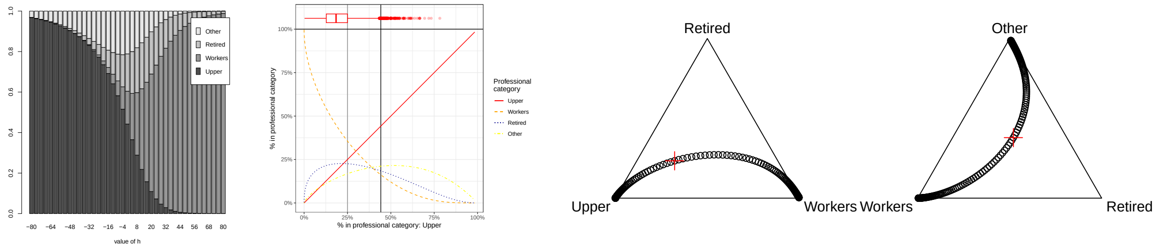

Figure 2: Change in a general direction

In particular, another direction of interest is the direction given by $\mathbf{u} = \mathbf{e}_m \ominus \mathbf{e}_{m'}$ for two indices m and m' between 1 and D_X: it is the closure of the vector with ones except for position m where it is e and position m' where it is e^{-1}. Moving in this direction therefore means multiplying the m^{th} share by a factor and dividing the ${m'}^{th}$ share by the same factor and closing the obtained vector. This move looks alike the isotemporal substitution scheme of [Dumuid et al., 2019]. However their absolute compensation scheme cannot be extended into a linear map with respect to the simplex operations. Our approach can viewed as the discrete version of [Jašková et al., 2023] who present a relative compensation scheme in a continuous CoDa framework. We will use this type of direction later on in Section 8.

5 Conditional plots for CoDa regression

A first approach to assess the effect of a given explanatory variable is to construct conditional plots. Using a predefined scenario of change of one explanatory variable, for a given observation that serves as an initial point, we can create a scatterplot of the predicted dependent variable versus the explanatory variable of interest conditional on the values of the other explanatory variables at the initial point. In a linear model, it is natural to define the change scenario along a linear path, and in the CoDa framework, we can use the simplex linear paths of the previous section. The following two subsections illustrate these conditional plots for our examples of a scalar dependent model explaining the turnout and a compositional dependent model explaining the vote shares of the candidates/blocks. The reference observation in all the conditional plots is always the municipality of Paris.

5.1 Variation of a scalar dependent with a compositional explanatory

As mentioned in Section 4, for a compositional explanatory variable, it is impossible to change a single component without changing the others due to the sum constraint. However, we may consider linear paths in the simplex as in Section 4 and plot the fitted shares evolution when the compositional variable evolves along this path. We distinguish two cases depending on whether the direction vector points to a vertex or not.

Figure 3 displays the conditional plots of the fitted turnout for the direction to the vertex “Upper” on the left and the direction as in Figure 2 on the right. The two bottom plots show the variations of the explanatory shares along the two paths as a function of h. On the y-axis of the two plots at the top show the corresponding prediction of the turnout. The two vertical lines on both graphs correspond to the predicted turnout of Paris (top) and the PC shares of Paris (bottom).

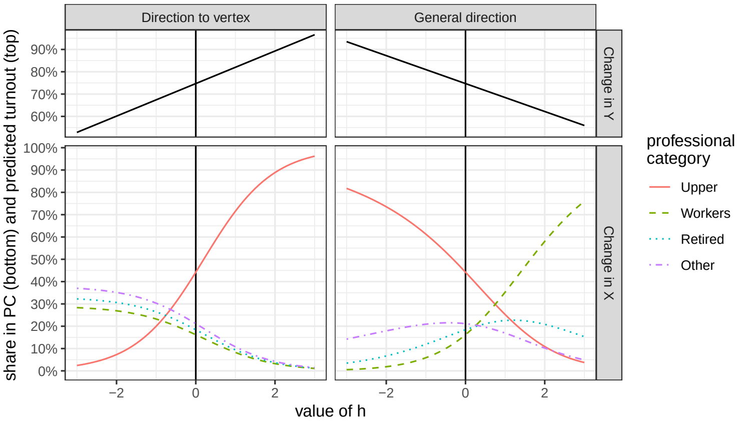

Figure 3: Variation of a scalar dependent as a function of a compositional explanatory

5.2 Variation of a compositional dependent

When the dependent variable is compositional, we need specific approaches for creating conditional plots of scalar and compositional dependent variables.

Let us first look at the conditional plot of a compositional dependent as a function of a scalar explanatory variable. Figure 4 displays the variations of the vote shares for each of the six candidates/blocks as a function of the population density. The vertical line indicates the actual population density in Paris, which is about 20000 inhabitants per km^2. Here we consider an extreme scenario where the population variation ranges from a tenfold increase to thousands of its original size to illustrate the behavior of the model.

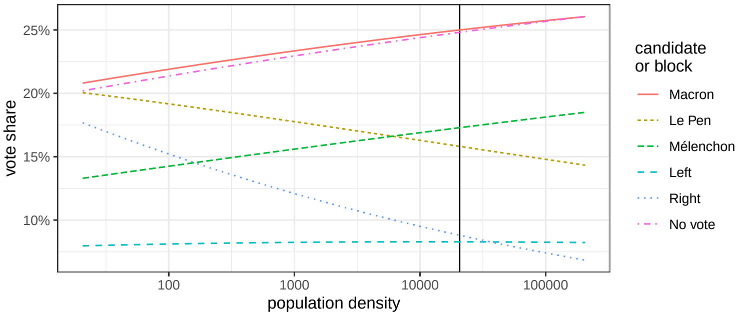

Figure 4: Variation of a compositional dependent as a function of a scalar explanatory variable

Turning now attention to the variations of a compositional dependent as a function of a compositional explanatory, Figure 5 displays two paths for Paris: the left graph corresponds to a change towards the “Upper” vertex and the right one to a general direction. The two bottom graphs are alternative representations of the same line path as the one illustrated by Figures 1 and 2. On the top left graph, we can see that moving in the direction of “Upper” almost does not affect the “Right” share, increases the shares of “Macron”, “Mélenchon” and “Left” and decreases the shares of “Le Pen” and “No vote”. The changes in the vote composition in the top right graph appear more complex, which is partially due to the non-monotonous evolution of the “Macron” and “No vote” shares. Jointly assessing the two right hand graphs further reveals that “Le Pen” is predicted to attract most of the votes when “Workers” dominate all the other professional categories.

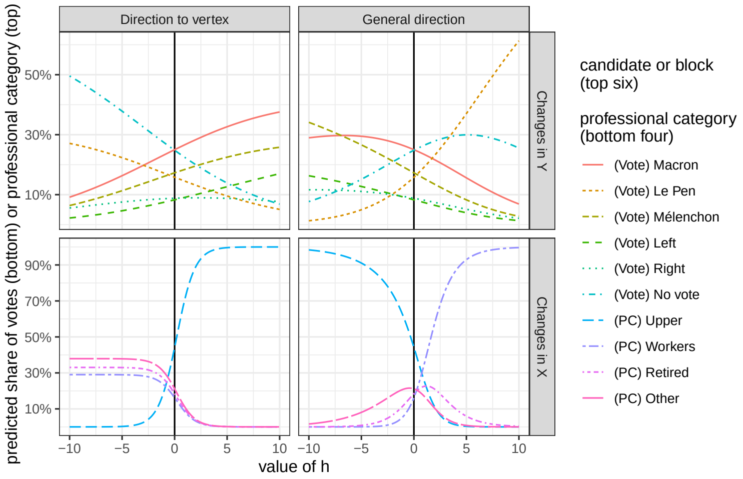

Figure 5: Variation of a compositional dependent as a function of a compositional explanatory

Another way of illustrating the conditional variations could be to plot as in Figures 6 and 7 two adjacent ternary diagrams showing the variations of a subcomposition of $\mathbf{X}$ along the path on the left and corresponding variations of a subcomposition of $\mathbf{Y}$ on the right. Additionally, color and shading can be linked to the value of h, indicating the correspondence between both ternary diagrams. In Figure 6, the path is towards the "Upper" vertex and in Figure 7 based on the general direction. However, while this type of illustration can be useful for low dimensional CoDa variables, it becomes cumbersome when the dimensions of the CoDa variables increase.

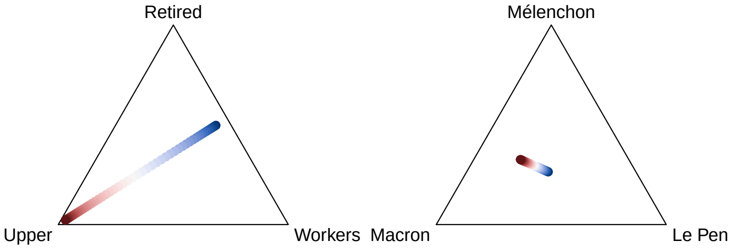

Figure 6: Variation of a compositional dependent as a function of a compositional explanatory with ternary diagrams for changes of $\mathbf{X}$ in the direction of a vertex

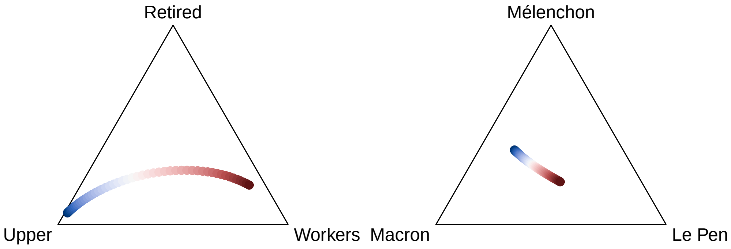

Figure 7: Variation of a compositional dependent as a function of a compositional explanatory with ternary diagrams for changes of $\mathbf{X}$ in a general direction

6 Impact of a CoDa explanatory on a scalar dependent

The following interpretations are similar to the classical ones like [Boogaart and Tolosana-Delgado, 2013], [van den Boogaart et al., 2021] or [Coenders and Pawlowsky-Glahn, 2020]. However, the present study of linear increments of Section 4 allows a generic treatment for all directions of changes in the explanatory space as opposed to the list of cases considered in [Coenders and Pawlowsky-Glahn, 2020] with which they are coherent.

As can be found in [Boogaart and Tolosana-Delgado, 2013], in model (8), the linear increment of Y (in real space) resulting from changing $\mathbf{X}$ with parameter $\boldsymbol{\beta}$ according to the path in (11) is

$$\tilde{Y}(h) - \tilde{Y}(0) = < \boldsymbol{\beta}, h \odot \boldsymbol{u} >_A = h < \boldsymbol{\beta}, \boldsymbol{u} >_A = h < \operatorname{clr} \boldsymbol{\beta}, \operatorname{clr} \boldsymbol{u} >_A . \tag{20}$$

The impacts are therefore completely described by the clr coordinates of the parameter $\boldsymbol{\beta}$ and are constant throughout observations. For directions of type $\boldsymbol{u} = \boldsymbol{e_m}$, the impact in (20) is driven by the observation independent semi-elasticity vector $\boldsymbol{SE}$ presented in [Morais and Thomas-Agnan, 2021]. This correspondence stems from the fact that in model (8) finite increments of size 1 coincide with infinitesimal increments (semi-log derivatives here) due to the linearity of Y with respect to $\operatorname{clr} \mathbf{X}$ in the classical sense.

Table 4: Impacts of change in PC on the turnout

	Direction to vertex	General direction
Initial value	0.747	0.747
New value	0.752	0.741
Semi elasticity	0.063	-0.075
Variation in % points	0.47 %	-0.56 %

Table 4 presents an example for the regression of the turnout on the explanatory variables of Table 2 based on linear increments in the professional categories shares in two directions: direction to the vertex "Upper" on the left as in Figure 1 and general direction as in Figure 2 on the right. The starting observation is Paris and the change is of size $h = 0.1$ corresponding to a relative increase of 5.58% of the share of the category "Upper" for the first direction. The first line of the table displays the initial turnout in Paris, the second line shows the value after the change of the professional categories. Also printed in the table are the values of the semi-elasticities and the absolute variation of Y measured in % points because it is a rate.

7 Infinitesimal impacts on a CoDa dependent

[Morais et al., 2018] and [Morais and Thomas-Agnan, 2021] introduce the use of semi-elasticities and elasticities to measure the changes on a CoDa dependent variable due to small variations of a scalar or a compositional variable, but restricting in the second case to the directions towards a vertex. After recalling the Taylor formulas which support these interpretations, we extend the second one to the case of a general direction. We then illustrate for our model of the election results how these approximations lead to interpretations in terms of relative variations of vote shares.

7.1 Taylor formulas

Thanks to Chapter 12 (written by Egozcue *et al.*) and Chapter 13 (written by Barcelo-Vidal et al.) of the book [Pawlowsky-Glahn et al., 2015] and to [Morais and Thomas-Agnan, 2021], we can write Taylor formulas to approximate functions to and/or from a simplex for small increments. Let us briefly summarize these results and complete them for the case of changes in a general direction.

Scalar argument and simplex value For a function $\mathbf{f} : \mathbb{R} \mapsto \mathcal{S}^D$, we find in Chapter 12 (written by Egozcue *et al.*) of the book [Pawlowsky-Glahn et al., 2015] the following approximation for a small change $x \longrightarrow x + h$ of x :

$$\begin{aligned}\mathbf{f}(x+h) &\simeq \mathcal{C}\left(f_1(x)\exp(h\frac{\partial \log f_1(x)}{\partial x}), \ldots, f_D(x)\exp(h\frac{\partial \log f_D(x)}{\partial x})\right) \\ &\simeq \left(f_1(x)(1+h\frac{\partial \log f_1(x)}{\partial x}), \ldots, f_D(x)(1+h\frac{\partial \log f_D(x)}{\partial x})\right). \qquad (21)\end{aligned}$$

The constraint $\sum_{m=1}^{D} f_m(x) = 1$ implies that $\sum_{m=1}^{D} f_m(x)\frac{\partial \log f_m(x)}{\partial x} = 0$ and hence the right hand side of (21) is indeed a CoDa vector. The product of the semi-elasticity by h is then the value of the corresponding relative change $\frac{f_m(x+h)-f_m(x)}{f_m(x)}$ of component m of the CoDa vector $\underline{\mathbf{f}}(x)$ induced by a small additive change h of x.

Simplex argument and simplex value For a function $\underline{\mathbf{f}} : \mathcal{S}^{D_X} \mapsto \mathcal{S}^{D_Y}$, arising from an homogeneous function of degree zero by $\mathbf{f}(\breve{\boldsymbol{x}}) = \underline{\mathbf{f}}(\mathbf{x})$ where $\mathbf{x} = \mathcal{C}(\breve{\mathbf{x}})$, [Morais and Thomas-Agnan, 2021] find the following Taylor approximation of $\underline{\mathbf{f}}(\mathbf{x})$ resulting from a small change $\mathbf{x} \longrightarrow \mathbf{x} \oplus \delta \odot \boldsymbol{e_m}$ of $\mathbf{x}$

$$\begin{aligned}\underline{\mathbf{f}}(\mathbf{x} \oplus h \odot \mathbf{e}_m) &= \mathcal{C}\left(f_1(\boldsymbol{x})\exp(h\frac{\partial \log f_1(\breve{\boldsymbol{x}})}{\partial \log \breve{x}_m}), \ldots, f_D(\breve{\boldsymbol{x}}))\exp(h\frac{\partial \log f_D(\breve{\boldsymbol{x}})}{\partial \log \breve{x}_m})\right) \\ &\simeq \left(f_1(\breve{\boldsymbol{x}})(1+h\frac{\partial \log f_1(\breve{\boldsymbol{x}})}{\partial \log \breve{x}_m}, \ldots, f_D(\breve{\boldsymbol{x}})(1+h\frac{\partial \log f_D(\breve{\boldsymbol{x}})}{\partial \log \breve{x}_m}\right). \qquad (22)\end{aligned}$$

The constraint $\sum_{j=1}^{D_Y} f_j(\breve{\boldsymbol{x}}) = 1$ implies that for all m, we get $\sum_{j=1}^{D_Y} f_j(\breve{\boldsymbol{x}})\frac{\partial \log f_j(\breve{\boldsymbol{x}})}{\partial \log \breve{x}_m} = 0$ showing that the right hand side of (22) is indeed a CoDa vector. The product of the elasticities by h is equal to the relative change $\frac{\underline{f}_j(\boldsymbol{x} \oplus h \odot \mathbf{e}_m) - \underline{f}_j(\boldsymbol{x})}{\underline{f}_j(\boldsymbol{x})}$ of the vector $\underline{\mathbf{f}}(\boldsymbol{x})$ induced by the linear increment of $\boldsymbol{x}$ in the direction of $\mathbf{e}_m$ with signed intensity h.

Let us now generalize this Taylor formula to an increment in the direction of a generic unit vector $\boldsymbol{u} \in \mathcal{S}^D$.

Theorem 7.1. *For a function $\underline{\mathbf{f}}$ from the simplex $\mathcal{S}^{D_X}$ to the simplex $\mathcal{S}^{D_Y}$, we have the following first order Taylor expansion of the components of $\underline{\mathbf{f}}(\mathbf{x} \oplus h \odot \mathbf{u})$ in the neighborhood of $h = 0$:*

$$\underline{f}_j(\boldsymbol{x} \oplus h \odot \mathbf{u}) \simeq \underline{f}_j(\boldsymbol{x})(1 + h\sum_{m=1}^{D_X} \frac{\partial \log f_j(\breve{\boldsymbol{x}})}{\partial \log \breve{x}_m} \log(u_m)), \qquad (23)$$

where u_m is the m^{th} component of $\mathbf{u}$, $m = 1, \ldots, D_X$.

As in the vertex direction case, one can check that the right hand side of (23) is a composition: indeed taking the derivative of the constraint $\sum_{j=1}^{D_Y} f_j(\breve{\boldsymbol{x}}) = 1$ with respect to $\log \breve{x}_m$ yields for any m

$$\sum_{j=1}^{D_Y} f_j(\breve{\boldsymbol{x}})\frac{\partial \log f_j(\breve{\boldsymbol{x}})}{\partial \log \breve{x}_m} \mid_{h=0} = 0. \qquad (24)$$

7.2 Absolute and relative variations of shares for a given observation

In this section, we apply formulas (21) and (23) to approximate, in the compositional dependent model (7), the changes of the conditional expectation $\tilde{\mathbf{Y}}_i(x)$ when a given explanatory variable $\mathbf{X}$ changes along a linear path. Let us recall that there are easy expressions to compute the semi-elasticities and elasticities from model parameters estimates in [Morais and Thomas-Agnan, 2021].

Evaluating the semi-elasticities in (21) at the observation points X_i, we get a semi-elasticity D_Y-vector $\mathbf{SE}_i$ with elements $\frac{\partial \log f_j}{\partial \log \check{x}_m}$ evaluated at the observation point i. Similarly, evaluating the elasticities in (23) at the observation points X_i, we get an elasticity $D_Y \times D_X$ matrix $\mathbf{E}_i$ with elements $\frac{\partial \log f_j}{\partial \log \check{x}_m}$ evaluated at the observation point i. It is easy to show that the row sums of this observation dependent elasticity matrix, corresponding to the sum of elasticities of one component of $\mathbf{Y}$ with respect to all components of $\mathbf{X}$, is zero.

For interpreting the impact of a scalar explanatory variable X on a compositional dependent variable $\mathbf{Y}$, formula (21) allows to evaluate the percent variation of each dependent share Y_i due to an additive change of X, using the semi-elasticities of $\mathbf{Y}$ with respect to X. Table 5 illustrates this for the model of Table 3. In our example, we increase $\log(POP\_DENSITY)$ by $h = 0.1$, which is approximately a 10% increase in the population. The first line reflects the initial shares of the VOTE composition in Paris and the second line shows the new composition after the change. In the third line, we have the semi-elasticity which corresponds to the logarithmic derivative in (21). The variation in % is computed from (21) and can be approximated by h times the semi-elasticities of VOTE with respect to $\log(POP\_DENSITY)$. To obtain the variation in percentage points, it suffices to subtract the old shares from the new ones. Finally, we get the variation in units by multiplying the percentage point variation with the total of the response variable, in our case the number of registered voters in Paris.

Table 5: Increase of the population density by 10%

	Macron	Le Pen	Mélenchon	Left	Right	No vote
Initial shares	0.2500	0.1582	0.1729	0.0828	0.0880	0.2481
New shares	0.2505	0.1575	0.1734	0.0828	0.0871	0.2487
Semi elasticity	0.0198	-0.0411	0.0313	-0.0011	-0.1077	0.0230
Variation in %	0.2 %	-0.41 %	0.31 %	-0.01 %	-1.08 %	0.23 %
Variation in % points	0.05 %	-0.06 %	0.05 %	0 %	-0.09 %	0.06 %
Variation in units	678	-889	739	-13	-1297	781

For interpreting the impact of a compositional explanatory on a compositional dependent, formula (22) (resp: formula (23)) allows to evaluate the percent variation of each dependent share Y_i due to a given linear change of the compositional variable in the direction of a vertex (resp: in a general direction).

Table 6 illustrates these changes for the case when the professional category composition is moved by $h = 0.1$ in the direction of the vertex "Upper". For this particular direction, we can use (14) to calculate the relative changes in the dependent composition, which correspond to an $\alpha_m = 5.58\%$ increase for the share of "Upper" and a 4.42% decrease for the three remaining PC shares in our example.

Table 6: Change of PC in the direction of the vertex "Upper"

	Macron	Le Pen	Mélenchon	Left	Right	No vote
Initial shares	0.250	0.158	0.173	0.083	0.088	0.248
New shares	0.251	0.157	0.174	0.084	0.088	0.246
Elasticity	0.060	-0.074	0.059	0.087	0.013	-0.087
Variation in %	0.6 %	-0.74 %	0.59 %	0.87 %	0.12 %	-0.87 %
Variation in % points	0.15 %	-0.12 %	0.1 %	0.07 %	0.01 %	-0.22 %
Variation in units	2038	-1601	1395	981	151	-2964

Table 7 illustrates the corresponding variations, for a change in a more general direction. Here we take the same direction as the one used to illustrate the linear increments in Figure 2. Since the relative changes in the PC composition are different for each component of $\mathbf{X}$ in this case, we do not report the α values.

Table 7: Change of PC in a general direction

	Macron	Le Pen	Mélenchon	Left	Right	No vote
Initial parts	0.250	0.158	0.173	0.083	0.088	0.248
New parts	0.248	0.162	0.171	0.082	0.087	0.250
Elasticity	-0.072	0.246	-0.136	-0.136	-0.088	0.087
Variation in %	-0.72 %	2.46 %	-1.36 %	-1.36 %	-0.88 %	0.87 %
Variation in % points	-0.18 %	0.39 %	-0.23 %	-0.11 %	-0.08 %	0.22 %
Variation in units	-2459	5325	-3215	-1537	-1056	2941

7.3 Elasticity distribution and theoretical bounds

[Morais and Thomas-Agnan, 2021] show that the elasticities and semi-elasticities are observation dependent through the expected vector of shares $\mathbb{E}^{\oplus}\mathbf{Y} \mid \mathbf{X}$ (conditional on all explanatory variables). However, we obtain observation-independent quantities for

$$\mathbf{G}_{D_Y}\mathbf{SE}_i = \mathrm{clr}(\boldsymbol{\gamma})$$
$$\mathbf{G}_{D_Y}\mathbf{E}_i = \mathbf{B}.$$

The above equations imply that the parameter matrix $\mathbf{B}$ corresponds to the difference between the elasticities and their corresponding row average. Similarly, $\mathrm{clr}(\boldsymbol{\gamma})$ is a difference between a semi-elasticity and the average over components of $\mathbf{Y}$ of the semi-elasticities. This also shows that the (semi-) elasticities are only observation dependent through this average since we have:

$$\mathbf{SE}_i = \mathrm{clr}(\boldsymbol{\gamma}) + \mathbf{1}_{D_Y}\overline{SE_i}. \tag{25}$$

Using the result $\mathbf{SE}_i = \mathrm{clr}(\boldsymbol{\gamma}) - \mathbf{1}_{D_Y}\tilde{\mathbf{Y}}_i\,\mathrm{clr}(\boldsymbol{\gamma})'$ of [Morais and Thomas-Agnan, 2021], where $\mathbf{1}_{D_Y}$ is a D_Y vector of ones, we may further deduce that $\overline{SE_i} = -\tilde{\mathbf{Y}}_i'\,\mathrm{clr}(\boldsymbol{\gamma})$. Since the fitted value $\tilde{\mathbf{Y}}_i$ is an element of $\mathcal{S}^{D_Y}$, we can derive bounds on the average semi-elasticity for each observation

$$\min(\mathrm{clr}(\boldsymbol{\gamma})) \leq -\overline{SE_i} \leq \max(\mathrm{clr}(\boldsymbol{\gamma})). \tag{26}$$

These bounds are easily converted into bounds on the $\mathbf{SE}_i$ by (25). Additionally, because $\mathrm{clr}(\boldsymbol{\gamma})$ is centered we know that for a balanced composition $\tilde{\mathbf{Y}}_i = \mathcal{C}\mathbf{1}_{D_Y}$ the $\overline{SE_i} = 0$ and $\mathbf{SE}_i = \mathrm{clr}(\boldsymbol{\gamma})$. In other words, the more balanced the fitted shares for a given observation the more its semi-elasticities resembles the clr-transformed parameters.

Figure 8 illustrates these insights for the parameters associated with the logarithm of the population density in the model of Table 3. The first six entries on the Y-axis correspond to the choices available to the voters and the X-axis displays the value of the corresponding clr parameter coordinate, as well as their confidence intervals (see Annex). At the bottom, we display the distribution of $-\overline{SE_i}$ over all municipalities. We clearly see that these values are contained in the bounds (26). Furthermore, equation (25) implies that we may obtain the distribution of the semi-elasticity for any component of the dependent variable by shifting the distribution of $-\overline{SE_i}$ by the corresponding component of the clr parameter vector.

Similar considerations apply to the elasticities with the difference that, with the dependent X being multivariate, we need to focus on one column $E_{i,\bullet m}$ of the elasticity matrix $\mathbf{E}_i$ at a time, corresponding to component m of $\mathbf{X}$ and all components of $\mathbf{Y}$. For the elasticities of $\tilde{\mathbf{Y}}_i$ with respect to the m^{th} component of $\mathbf{X}$ we have

$$E_{i,\bullet m} = B_{\bullet m} + \mathbf{1}_{D_Y}\overline{E_{i,\bullet m}}. \tag{27}$$

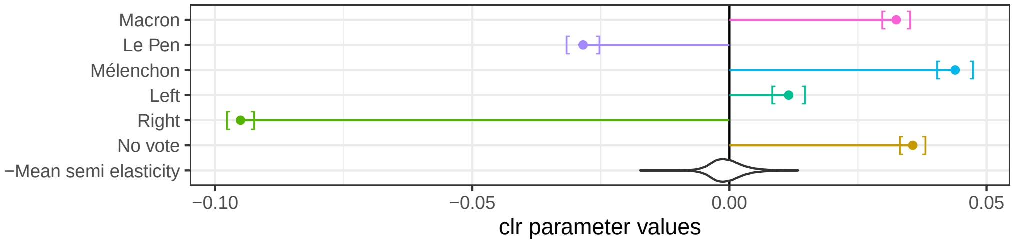

Figure 8: Semi elasticities distribution and confidence intervals for the clr parameters

Using the equation $\mathbf{E}_i = \mathbf{B} - \mathbf{1}_{D_Y} \tilde{\mathbf{Y}}_i \mathbf{B}$ from [Morais and Thomas-Agnan, 2021] we can derive that $\overline{E_{i,\bullet m}} = -\mathbf{1}_{D_Y} \tilde{\mathbf{Y}}_i B_{\bullet m}$, which leads to the interval:

$$\min(B_{\bullet m}) \leq \overline{E_{i,\bullet m}} \leq \max(B_{\bullet m}). \tag{28}$$

In Figure 9, we illustrate these results for the model of Table 3 where the parameter matrix $\mathbf{B}$ is associated with the compositions of the professional categories. Each facet in the graphic corresponds to one professional category and illustrates one column of the matrix $\mathbf{B}$. As before, the last value on the Y-axis shows the distribution of the negative mean elasticity, which is contained in the interval defined by the maximal and minimal parameter values in the given column of $\mathbf{B}$.

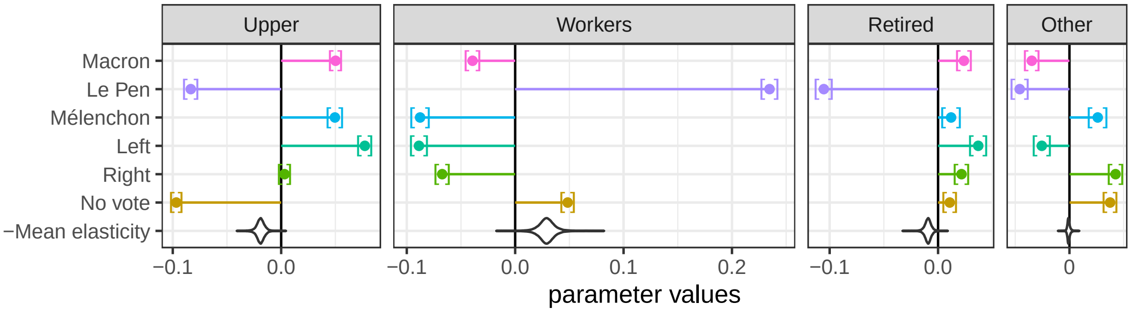

Figure 9: Elasticities distribution and confidence interval for the clr parameters

8 Relative variation of share ratios

We have seen that one limitation of the semi-elasticities/elasticities approach is that in the compositional dependent case these quantities are observation dependent. Moreover, they depend on the predicted value of the share vector $\mathbf{Y}$, making them non-linear functions of the parameters which means that we cannot easily evaluate their variability. We are going to overcome this limitation by taking advantage of the linearity of the model with respect to the simplex operations. Indeed since model (7) is linear for the perturbation and powering operations, a linear change of any of the explanatory variables will result in a linear change in the expected dependent variable, linearity being understood with respect to the geometry of each considered variable. By using (15) and (16) of Section 4, we have learned that a linear change of any simplex variable can be written as a finite multiplicative change of its share ratios, which can be approximated for small h by a percent variation of its share ratios. It is therefore natural to see share ratios appearing in the ensuing interpretation as we detail below. The results of this section are coherent with the share ratio interpretation found in [Wang et al., 2013] in a restricted model (see

[Dargel and Thomas-Agnan, 2023] for understanding the peculiarity of this model). We will use the notation $\tilde{\mathbf{Y}}(x)$ of Sections 3.2 and 3.3 to express the impacts of the explanatory variables on the expected value of the dependent variable.

For a compositional variable $\mathbf{X}$, taking the logarithmic derivative of (15) yields, for any couple of indices j and l between 1 and D_X,

$$\frac{d}{dh} \log \frac{X_j(h)}{X_l(h)} = \log \frac{u_j}{u_l}. \tag{29}$$

This result will be used on the right hand side and/or left hand side of the regression equation depending on how compositional variables are used in the model.

8.1 Model with a scalar dependent

In model (8), from (20) and (29), it is easy to prove that, when $\mathbf{X}$ moves along (11), we have the following absolute variation of $\tilde{Y}(h)$ with respect to the relative variations of the share ratios of $\mathbf{X}$ (for $j \neq l$ and whenever $u_j \neq u_l$)

$$\frac{\frac{d\tilde{Y}(h)}{dh}}{\frac{d}{dh} \log \frac{X_j}{X_l}} = \frac{< \mathrm{clr}\, \boldsymbol{\beta}, \mathrm{clr}\, \boldsymbol{u} >_A}{\log \frac{u_j}{u_l}} = \frac{SE_j \log(u_j) + SE_l \log(u_l)}{\log(u_j) - \log(u_l)} \tag{30}$$

Remembering that $\mathrm{clr}\, \boldsymbol{\beta}$ corresponds to the vector of semi-elasticicities of Y with respect to all shares of $\mathbf{X}$, we see that the absolute variations of Y generated by relative changes in the share ratios of $\mathbf{X}$. are driven by a simple linear combination of two semi-elasticities. An interesting consequence of (30) can be derived:

$$SE_j - SE_l = \mathrm{clr}(\beta_j) - \mathrm{clr}(\beta_l) = \log(\beta_j/\beta_l). \tag{31}$$

Note that (31) can alternatively be obtained using the expressions of the semi-elasticities as a function of the parameters. Equation (31) shows that the ratios of $\mathbf{X}$ which have the highest impact on Y are those ratios with the highest values of the corresponding ratio of the $\boldsymbol{\beta}$ components, making the semi-elasticity differences a privileged tool for summarizing the data. Since the estimates of $\boldsymbol{\beta}$ are usually obtained in ilr or alr coordinates, we refer to the transition formulas in the Annex to derive their clr counterparts and the associated variances. The confidence intervals for the differences of $\mathrm{clr}\, \boldsymbol{\beta}$ components can be computed easily, and therefore we can determine which ratios of $\mathbf{X}$ significantly impact Y.

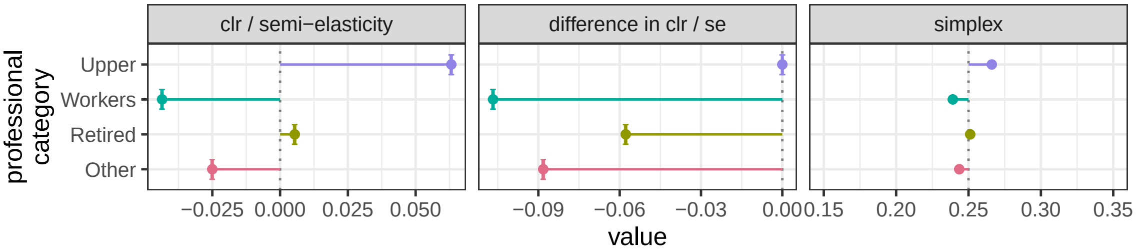

Figure 10: Three representations for the impact of professional categories

Figure 10 depicts, in three ways, the impact of the professional categories in our model explaining the election turnout rate. The left graphic illustrates the clr parameters and their confidence intervals, which are extremely narrow due to the large number of observations in our example. As mentioned previously, these clr values coincide with the semi-elasticities. The middle graph shows confidence intervals and point estimates of the semi-elasticity differences, where the share of "Upper" is used in the denominator (indexed by l in (31)). Finally, the right graph depicts the parameter values in the simplex, which can be useful as an impression of their magnitude. In our model, $\boldsymbol{\beta}$ is very close to the balanced composition that plays the role of the zero-vector in the simplex. Thus, we may conclude that only strong imbalances in the professional category composition can lead to large changes in the turnout rate.

8.2 Model with a compositional dependent

When $\mathbf{Y}$ is compositional, the conditional expectation $\tilde{\mathbf{Y}}(x)$ is also a function of h: along the path (11) for compositional exploratory variables and along the path $Z(h) = Z(0) + h$ for scalar ones. In both cases, we will denote it by $\tilde{\mathbf{Y}}(h)$.

8.2.1 Scalar explanatory

In model (7), let us first consider the impact of a scalar explanatory variable Z with parameter γ for which we have $\tilde{\mathbf{Y}}(h) = \tilde{\mathbf{Y}}(0) \oplus h \odot \boldsymbol{\gamma}$.

Using the same technique as in Section 4 to go from equation (11) to the share ratio finite increment (15), we can derive the following share ratio finite increment result, for observation unit i and for any two components j and l between 1 and D_Y :

$$\frac{\tilde{\mathbf{Y}}_{ij}(h)}{\tilde{\mathbf{Y}}_{il}(h)} = \frac{\tilde{\mathbf{Y}}_{ij}(0)}{\tilde{\mathbf{Y}}_{il}(0)} \exp(h \log \frac{\gamma_j}{\gamma_l}) \tag{32}$$

Taking the log and then the derivative of (32) yields the share ratio infinitesimal relative change

$$\frac{\frac{d}{dh} \log \frac{\tilde{Y}_{ij}(h)}{\tilde{Y}_{il}(h)}}{\frac{d}{dh} Z} = \log \frac{\gamma_j}{\gamma_l} = SE_{ij} - SE_{il}, \tag{33}$$

where the difference $SE_{ij} - SE_{il}$ is observation independent, although each term by itself is not.

Figure 11 displays these differences for the model of Table 3, where the vote share of Macron is in the denominator (l in the formula). The vote share whose elasticity is most different from that of Macron is that of the "Right" block. In contrast, the share of "No votes" reacts not significantly differently than the share of Macron to a change in the population density.

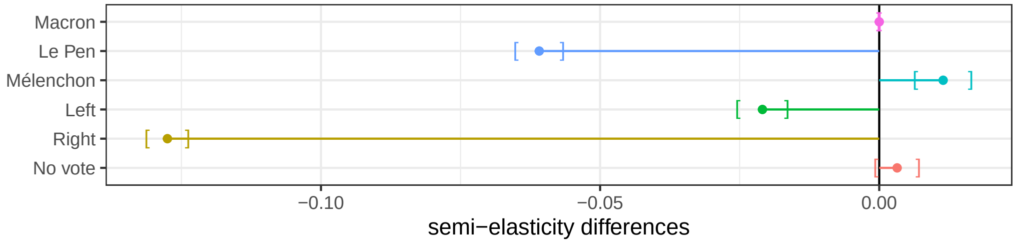

Figure 11: Confidence interval for the differences in semi-elasticities parameters

8.2.2 Compositional explanatory

Similarly, in model (7), for evaluating the impact of a compositional covariate $\mathbf{X}$ with matrix of parameters $\mathbf{B}$ on the compositional dependent $\mathbf{Y}$, let us express the simplex variations of the explanatory variable and the dependent variable as a function of h when $\mathbf{X}$ changes along the linear path in (11).

As in (29), since $\tilde{\mathbf{Y}}(h) = \tilde{\mathbf{Y}}(0) \oplus \mathbf{B} \boxdot \mathbf{u}$, if we let $\mathbf{u_B} = \mathbf{B} \boxdot \mathbf{u}$, we get easily for any observation i and any couple of indices j and l between 1 and D_Y

$$\frac{d}{dh} \log \frac{\tilde{Y}_{ij}(h)}{\tilde{Y}_{il}(h)} = \log \frac{u_{Bj}}{u_{Bl}} = \sum_{k=1}^{D_X} (B_{jk} - B_{lk}) \log u_k. \tag{34}$$

We can conclude that when $\mathbf{X}$ moves along the path (11), the relative change of the dependent share ratio $\frac{\tilde{Y}_j(h)}{\tilde{Y}_l(h)}$ with respect to the explanatory share ratio $\frac{\tilde{X}_{j'}(h)}{\tilde{X}_{l'}(h)}$, which we will call a share ratio elasticity, is observation independent and is given by

$$\frac{\frac{d}{dh}\log\frac{\tilde{Y}_{ij}(h)}{\tilde{Y}_{il}(h)}}{\frac{d}{dh}\log\frac{\tilde{X}_{ij'}(h)}{\tilde{X}_{il'}(h)}} = \frac{\sum_{k=1}^{D_X}(B_{jk}-B_{lk})\log u_k}{\log\frac{u_{j'}}{u_{l'}}}. \tag{35}$$

It is obvious that this share ratio elasticity is generally only meaningful when both share ratios involve distinct components ($j' \neq l'$ and $j \neq l$), when $u_{j'} \neq u_{l'}$ and $u_{j'}, u_{l'} \neq 0$. This latter condition is generally satisfied for valid directions, i.e. that contain no zero components.

Let us now look at the form of (35) for the particular directions given by $\mathbf{e}_{j'}$ and by $\mathbf{e}_{j'} \ominus \mathbf{e}_{l'}$. For the direction $\boldsymbol{u} = \mathbf{e}_{j'}$, the whole expression in (35) becomes

$$\sum_{k=1}^{D_X}(B_{jk}-B_{lk})\log u_k = B_{jj'} - B_{lj'} = E_{i,jj'} - E_{i,lj'}, \tag{36}$$

simplifying to an observation independent elasticity difference, although one should note that the individual elasticity terms depend on the observation i. Figure 12 shows these elasticity differences, where each facet corresponds to one of the four professional category components (j'). The six elements of the Y-axis correspond to the numerator of the dependent share ratios (j), where Macron is again used in the denominator (l).

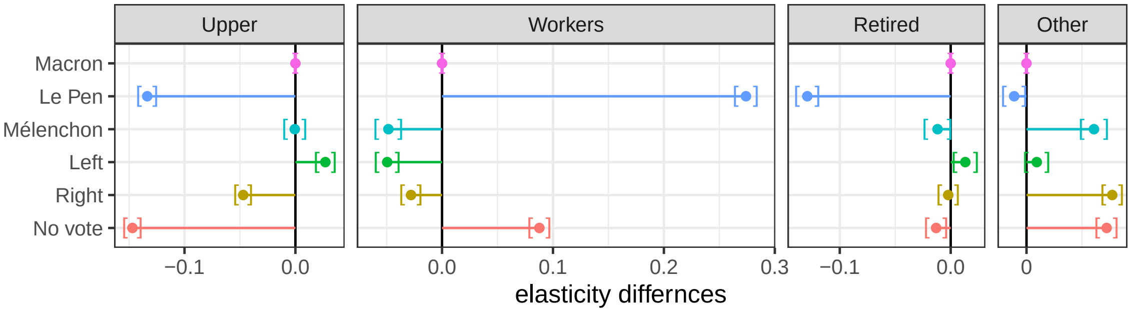

Figure 12: Confidence interval for the differences in elasticities parameters

For the direction $\mathbf{e}_{j'} \ominus \mathbf{e}_{l'}$ the denominator of (35) equals two and the numerator becomes

$$\begin{aligned}\sum_{k=1}^{D_X}(B_{jk}-B_{lk})\log u_k &= B_{jj'} - B_{lj'} - B_{jl'} + B_{ll'} \\ &= E_{i,jj'} - E_{i,lj'} - E_{i,jl'} + E_{i,ll'},\end{aligned} \tag{37}$$

which is a double difference in elasticities. While this may appear more complex than (36), it has the advantage that all changes in the dependent shares are exclusively driven by changes in the two components in the $\mathbf{X}$ share ratio, as can be verified by looking at the difference vector $\operatorname{clr}\mathbf{X}(0) - \operatorname{clr}\mathbf{X}(h)$, for the direction $\mathbf{e}_{j'} \ominus \mathbf{e}_{l'}$. As is explained in the following, this double difference also enables a more nuanced view of how pairwise trade-offs between components of $\mathbf{X}$ affect the share-ratios of $\mathbf{Y}$.

To exploit the share ratio elasticities in our model explaining the vote shares, let us first be clear about its dimensions. In this model, the vote share composition of dimension $D_Y = 6$ is explained by the professional categories composition with $D_X = 4$, which implies that (35) leads to $D_X^2 D_Y^2 = 576$

potential combinations of $l, j = 1, ..., D_Y$ and $l', j' = 1, ..., D_X$. However, since share ratios based on the same components ($l = j$ or $l' = j'$) are meaningless, the number of informative values drops to 360. We could thus represent these share ratio elasticities in a 12 × 30 matrix, with columns referring to share ratios of Y and rows referring to share ratios of X. Additionally, this matrix may be grouped into blocks based on the shares $X_{l'}$ and Y_l in the denominators, leading to $D_Y D_X = 24$ blocks, each containing 15 elasticity values.

Figure 13 shows this share ratio elasticity matrix for the relative compensation directions of the form $\boldsymbol{u} = \mathbf{e}_{j'} \ominus \mathbf{e}_{l'}$, where j' refers to the numerator and l' to the denominator in each row of the figure. The block second from the top and second from the right is the most striking, in that it only contains positive values with relatively large magnitudes. Based on the share ratios on the X and Y axis, we may interpret the values in this block as follows: if any of the professional categories grows at the expense of "Workers," all candidates gain votes at the expense of Le Pen. This effect is, however, weaker for "No vote" than for the other candidates. In contrast, if we now concentrate on the next block to the right, we see that, when any professional category grows at the expense of "Workers", all other candidates either lose votes in favor of Mélenchon or are almost unaffected. The candidate that would lose the most is Le Pen followed by "No vote". These findings indicate that individuals in the professional category "Workers" predominantly prefer to vote for Le Pen or not to cast a (valid) vote over voting for Mélenchon.

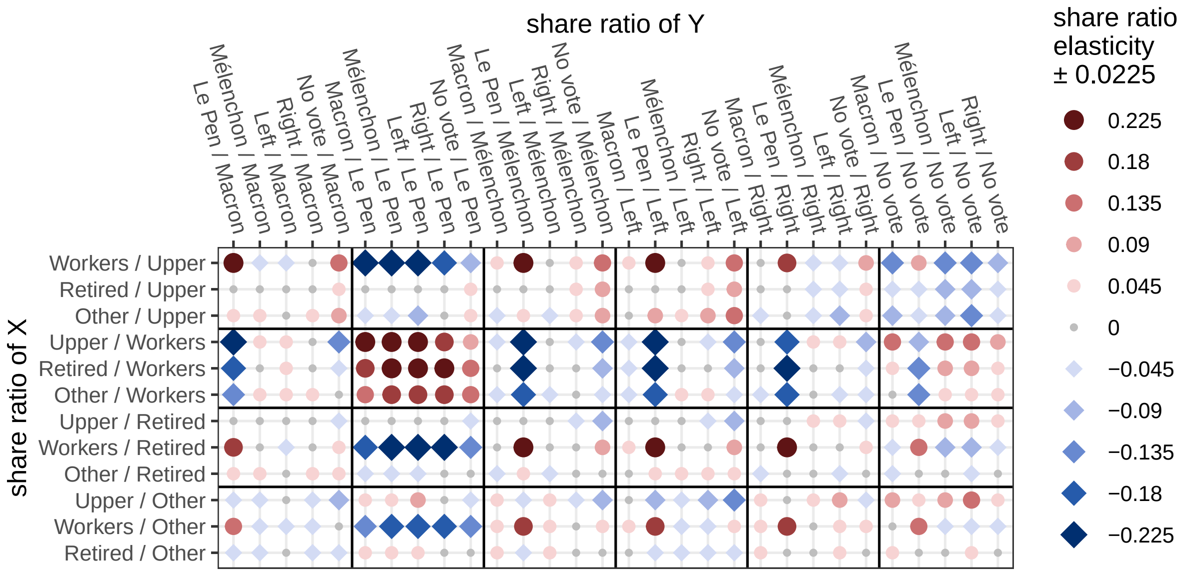

Figure 13: Share ratio elasticities for relative compensation directions

Another way to make sense of Figure 13 is to consider the depicted values in each block as a decomposition of the parameter matrix. In our example, the dimension of $\mathbf{B}'$ (4 × 6) matches the block structure of the share ratio elasticity matrix in Figure 13. We can further use (37) and the fact that $\mathbf{B}$'s rows and columns sum to zero to derive the sum of all elements in each block indexed by (l, l') as

$$\frac{1}{2} \sum_{j \neq l}^{D_Y} \sum_{j' \neq l'}^{D_X} B_{jj'} - B_{lj'} - B_{jl'} + B_{ll'} = \frac{1}{2} D_X D_Y B_{ll'}. \tag{38}$$

The above equation proves that the share ratio elasticities are indeed, up to a proportionality constant, a decomposition of the elements of $\mathbf{B}$. Put differently, the totals in each block of the share elasticity matrix are summarized by $\frac{1}{2} D_X D_Y \mathbf{B}'$. The image of the matrix $\mathbf{B}'$ in Figure 14 confirms this link visually. The inspection of the **B** matrix of this model already reveals that the strongest impacts on **Y** in this example occur for share ratios involving "Le Pen" and "Workers" in their denominators.

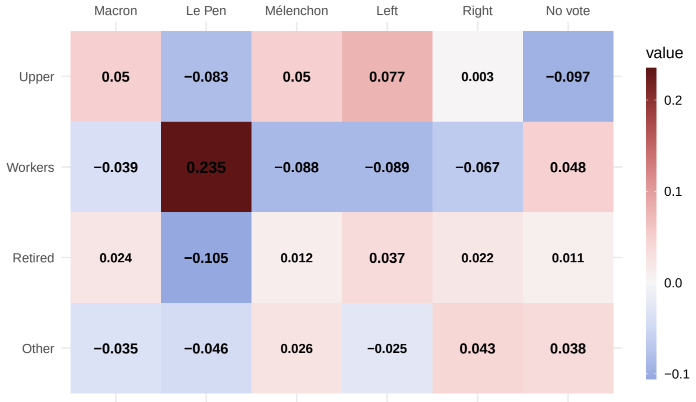

Figure 14: Image of the transposed parameter matrix

9 Conclusion

After briefly reviewing the existing approaches for interpreting the impact of a covariate in a CoDa regression model, we propose new approaches based on share ratios. These interpretations, as well as those based on semi-elasticities or elasticities, allow a direct interpretation in the space of shares (the simplex) rather than in a coordinate space (which depend on some transformation). In order to derive them, we establish some simple properties of linear increments in the simplex and a new Taylor formula for approximating a simplex valued function of a simplex variable moving in a general direction. We then show that, in these models, infinitesimal share-ratio variations are independent from the observation unit, making them natural tools for interpretations. We also adapt conditional plots to the framework of CoDa regression models. A real data example about the French presidential elections in 2022 is used to illustrate conditional plots, elasticities, semi-elasticities and share ratio elasticities. All the presented tools are available in the R package **CoDaImpact**, which is intended to be used in coordination with the R package **compositions**, and illustrated with a vignette [3].

10 Annex

Transition formulas between log-ratio spaces

This section presents transition formulas between the three families of log-ratio spaces, usually indexed by their corresponding transformations by alr, clr, and ilr. One interesting use of these formulas is the conversion of estimated parameters from one space to another without having to re-estimate the model. For example, it is often more practical to estimate a CoDa model in an ilr space, while the model interpretations in terms of elasticities are directly linked to the clr parameters.

[3] `https://github.com/LukeCe/CoDaImpact`

The log-ratio transformations for $\boldsymbol{x} \in \mathcal{S}^D$ and their corresponding inverses can be expressed as:

$$\begin{aligned} \operatorname{alr}_j(\boldsymbol{x}) &= \mathbf{F}_j \log(\boldsymbol{x}) & \operatorname{alr}^{-1}(\boldsymbol{x}) &= \mathcal{C}(\exp(\mathbf{K}_j \mathbf{x}^*)) \\ \operatorname{clr}(\boldsymbol{x}) &= \mathbf{G}_D \log(\boldsymbol{x}) & \operatorname{clr}^{-1}(\boldsymbol{x}) &= \mathcal{C}\left(\exp(\mathbf{x}^*)\right) \\ \operatorname{ilr}_V(\boldsymbol{x}) &= \mathbf{V}' \log(\boldsymbol{x}) & \operatorname{ilr}^{-1}(\boldsymbol{x}) &= \mathcal{C}\left(\exp(\mathbf{V}\mathbf{x}^*)\right), \end{aligned}$$

where the matrix $\mathbf{F}_j$ $(D-1 \times D)$ can be derived from $\mathbf{I}_{D-1}$ by inserting the vector $-\mathbf{1}_{D-1}$ as j^{th} column and $\mathbf{K}_j$ $(D \times D-1)$ is obtained from $\mathbf{G}_D$ by removing its j^{th} column. These matrices are further linked by the following identities:

$$\begin{aligned} \mathbf{K}_j \mathbf{F}_j &= \mathbf{G}_D & \mathbf{F}_j \mathbf{K}_j &= \mathbf{I}_{D-1} \\ \mathbf{V}\mathbf{V}' &= \mathbf{G}_D & \mathbf{V}'\mathbf{V} &= \mathbf{I}_{D-1}. \end{aligned}$$

It is then easy to derive the following transition formulas that convert vectors between the different log-ratio spaces:

$$\begin{aligned} \operatorname{clr}(\boldsymbol{x}) &= \mathbf{K}_j \operatorname{alr}_j(\boldsymbol{x}) & \operatorname{alr}_j(\boldsymbol{x}) &= \mathbf{F}_j \ \operatorname{clr}(\boldsymbol{x}) \\ \operatorname{clr}(\boldsymbol{x}) &= \mathbf{V} \operatorname{ilr}_V(\boldsymbol{x}) & \operatorname{ilr}_V(\boldsymbol{x}) &= \mathbf{V}' \operatorname{clr}(\mathrm{x}) \end{aligned}$$

Similar formulas can be found in the case matrix versions of the log-ratio spaces. This is particularly interesting for the YX compositional model where the influence of a compositional variable $\boldsymbol{x} \in \mathcal{S}^{D_X}$ on a compositional dependent variable $\boldsymbol{y} \in \mathcal{S}^{D_Y}$ is mediated by the matrix $\mathbf{B} \in \mathcal{A}_{D_Y D_X}$. We will use the star notation $\mathbf{B}^*$ for the matrix in the transformed space. Let us first note that $\mathbf{B}^* = \mathbf{G}_{D_Y} B \mathbf{G}_{D_Y} = \mathbf{B}$ for the case when both variables are clr transformed ($\boldsymbol{x}^* = \operatorname{clr}(\boldsymbol{x})$ and $\boldsymbol{y}^* = \operatorname{clr}(\boldsymbol{y})$). The full list of cases is summarized in the table below.

Transformations		Link between $\mathbf{B}$ and $\mathbf{B}^*$		
$\boldsymbol{x}$	$\boldsymbol{y}$			
clr	clr	$\mathbf{B} = \mathbf{B}^*$		
	alr_l	$\mathbf{B} = \mathbf{K}_l^Y \mathbf{B}^*$	$\iff$	$\mathbf{B}^* = \mathbf{F}_l^Y \mathbf{B}$
	ilr_{V_Y}	$\mathbf{B} = \mathbf{V}_Y' \mathbf{B}^*$	$\iff$	$\mathbf{B}^* = \mathbf{V}_Y \mathbf{B}$
alr_j	clr	$\mathbf{B} = \mathbf{B}^* \mathbf{F}_j^X$	$\iff$	$\mathbf{B}^* = \mathbf{B}\mathbf{K}_j^X$
	alr_l	$\mathbf{B} = \mathbf{K}_l^Y \mathbf{B}^* \mathbf{F}_j^X$	$\iff$	$\mathbf{B}^* = \mathbf{F}_l^Y \mathbf{B}\mathbf{K}_j^X$
	ilr_{V_Y}	$\mathbf{B} = \mathbf{V}_Y' \mathbf{B}^* \mathbf{F}_j^X$	$\iff$	$\mathbf{B}^* = \mathbf{V}_Y \mathbf{B}\mathbf{K}_j^X$
ilr_{V_X}	clr	$\mathbf{B} = \mathbf{B}^* \mathbf{V}_X$	$\iff$	$\mathbf{B}^* = \mathbf{B}\mathbf{V}_X'$
	alr_j	$\mathbf{B} = \mathbf{K}_l^Y \mathbf{B}^* \mathbf{V}_X$	$\iff$	$\mathbf{B}^* = \mathbf{F}_l^Y \mathbf{B}\mathbf{V}_X'$
	ilr_{V_Y}	$\mathbf{B} = \mathbf{V}_Y' \mathbf{B}^* \mathbf{V}_X$	$\iff$	$\mathbf{B}^* = \mathbf{V}_Y \mathbf{B}\mathbf{V}_X'$

Variances of clr parameters and of elasticities differences

The transition formulas of the previous paragraph also allow to derive the variances of the parameters in any log ratio space. We exemplify this transition for the case of a model that is estimated in an ilr space, to obtain the variances of the clr parameters. In a second step, the clr variances are used to derive the variances of the differences in elasticities and semi-elasticities presented in Section 8.

Let us first consider a CoDa model as in (7) with a single compositional explanatory associated with the parameter matrix $\mathbf{B}$. The parameters can be estimated in an ilr space using the OLS estimator $\hat{\mathbf{B}}^* = (\mathbf{X}^{*\prime}\mathbf{X}^*)^{-1}(\mathbf{X}^*\mathbf{Y}^*)$, where $\mathbf{X}^* = \operatorname{ilr}_{V_X}(\mathbf{X})$, $\mathbf{Y}^* = \operatorname{ilr}_{V_Y}(\mathbf{Y})$ and $\hat{\mathbf{B}}^* = \mathbf{V}_Y{}'\hat{\mathbf{B}}\mathbf{V}_X$. A classical result of multivariate regression models then expresses the variance of the vectorized estimator as

$$\mathbb{V}\mathrm{ar}(\operatorname{Vec}\hat{\mathbf{B}}^*) = (\mathbf{X}^{*\prime}\mathbf{X}^*)^{-1} \otimes \hat{\Sigma}^*, \tag{39}$$

where the Vec operator stacks the columns of a matrix and $\otimes$ denotes the Kronecker product, and where the ilr covariance matrix $\hat{\Sigma}^* = \mathbf{V}_Y{}'\hat{\Sigma}\mathbf{V}_Y$ can be estimated directly from the ilr residuals. Exploiting the properties of the contrast matrices $\mathbf{V}_Y$ and $\mathbf{V}_X$, we can link the vectorized ilr estimator to its clr counterpart which is denoted $\hat{\mathbf{B}}$, without star in the following:

$$\mathrm{Vec}(\hat{\mathbf{B}}) = \mathrm{Vec}(\mathbf{V}_Y\hat{\mathbf{B}^*}\mathbf{V}_X{}') = (\mathbf{V}_X \otimes \mathbf{V}_Y)\,\mathrm{Vec}(\hat{\mathbf{B}^*}) \tag{40}$$

Inserting the above result in (39) and the expression of $\hat{\Sigma}^*$ then allows to derive the variance of $\hat{\mathbf{B}}$ as

$$\begin{aligned}\mathbb{V}\mathrm{ar}(\mathrm{Vec}\,\hat{\mathbf{B}}) &= (\mathbf{V}_X \otimes \mathbf{V}_Y)\,\mathbb{V}\mathrm{ar}(\mathrm{Vec}\,\hat{\mathbf{B}^*})(\mathbf{V}_X \otimes \mathbf{V}_Y)' \\ &= \mathbf{V}_X(\mathbf{X}^*{}'\mathbf{X}^*)^{-1}\mathbf{V}_X{}' \otimes \mathbf{G}_{D_Y}\hat{\Sigma}\mathbf{G}_{D_Y}.\end{aligned} \tag{41}$$

The steps to derive the conditional variances of parameter a vector γ associated with a scalar explanatory in model (7) are analogous but simpler since $\mathrm{clr}(\boldsymbol{\gamma})$ is a vector. In particular, we do not require Kronecker products or the Vec operator to derive

$$\mathbb{V}\mathrm{ar}(\mathrm{ilr}_{V_Y}\,\hat{\boldsymbol{\gamma}}) = (Z'Z)^{-1}\hat{\Sigma}^* \qquad \text{and} \qquad \mathbb{V}\mathrm{ar}(\mathrm{clr}\,\hat{\boldsymbol{\gamma}}) = (Z'Z)^{-1}\mathbf{G}_{D_Y}\hat{\Sigma}\mathbf{G}_{D_Y}.$$

In Section 8 we point out that the differences in elasticities and semi-elasticities coincide with the differences in clr parameters. Thus, to compute the variances of the (semi-) elasticities differences we can simply apply the classical formula for the variance of a sum of two random variables. In the case of semi-elasticity difference, for example, we have for any $j, l \in \{1, ..., D_Y\}$:

$$\mathbb{V}\mathrm{ar}(SE_j - SE_l) = \mathbb{V}\mathrm{ar}((\mathrm{clr}\,\hat{\boldsymbol{\gamma}})_j) + \mathbb{V}\mathrm{ar}((\mathrm{clr}\,\hat{\boldsymbol{\gamma}})_l) - 2\,\mathbb{C}\mathrm{ov}((\mathrm{clr}\,\hat{\boldsymbol{\gamma}})_j, (\mathrm{clr}\,\hat{\boldsymbol{\gamma}})_l).$$

Additional Proofs

Proof of (12) To demonstrate the result of (12) we focus on one component the equation $\boldsymbol{x}(h) = \boldsymbol{x}(0) \oplus h \odot \mathbf{u}$ and replace the simplex operators by their definition, allowing to derive

$$\begin{aligned} x_j(h) &= \frac{x_j(0)u_j^h}{\sum_{k=1}^D x_k(0)u_k^h} \\ \iff \frac{x_j(h)}{x_j(0)} &= \frac{u_j^h}{\sum_{k=1}^D x_k(0)u_k^h} \\ \iff \frac{x_j(h) - x_j(0)}{x_j(0)} &= \frac{u_j^h - \sum_{k=1}^D x_k(0)u_k^h}{\sum_{k=1}^D x_k(0)u_k^h}. \end{aligned} \tag{42}$$

Then, for small h, we can approximate $u_j^h \simeq 1 + h\log(u_j)$, leading to

$$\begin{aligned} \frac{x_j(h) - x_j(0)}{x_j(0)} &\simeq \frac{1 + h\log(u_j) - \sum_{k=1}^D x_k(0) - h\sum_{k=1}^D x_k(0)\log(u_k)}{\sum_{k=1}^D x_k(0) - h\sum_{k=1}^D x_k(0)\log(u_k)} \\ &\simeq \frac{h\log(u_j) - h\sum_{k=1}^D x_k(0)\log(u_k)}{1 - h\sum_{k=1}^D x_k(0)\log(u_k)} \\ &\simeq \frac{h\sum_{k=1}^D x_k(0)\log(\frac{u_j}{u_k})}{1 - h\sum_{k=1}^D x_k(0)\log(\frac{u_j}{u_k})}, \end{aligned} \tag{43}$$

where we exploit that $x_1(0) + \ldots + x_D(0) = 1$. The final expression in 43 lends itself to a series expression of the form $x/(1-x) = x + x^2 + x^3 + \ldots$. For small x the higher-order terms vanish, leading to

$$\frac{x_j(h) - x_j(0)}{x_j(0)} \simeq h\sum_{k=1}^D x_k(0)\log(\frac{u_j}{u_k}). \quad \square \tag{44}$$

□

Proof of (13) **and** (14) These results correspond to a special case of (44) for the direction direction $\boldsymbol{u} = \boldsymbol{e_m} = \mathcal{C}(1, \ldots, e, \ldots, 1)$, where e is at the m^{th} component. In this case, the relative share increments are defined in terms of the log-ratio $\log(e_{m,i}/e_{m,j})$. This ratio is not impacted by the closure operation used in the definition of $\boldsymbol{e_m}$, which means that we can apply equation (44) with $\boldsymbol{u} = (1, \ldots, e, \ldots, 1)$, leading to

$$\begin{aligned} \text{if} \qquad l \neq m, \quad \frac{z_l(h) - x_l(0)}{x_l(0)} \simeq & h \left(\sum_{k \neq m} z_m(0) \log \frac{1}{1} + z_m(0) \log \frac{1}{e} \right) \\ = & - h x_m(0) \quad \square \end{aligned}$$

$$\begin{aligned} \text{if} \qquad l = m, \quad \frac{z_l(h) - x_l(0)}{x_l(0)} \simeq & h \left(\sum_{k \neq m} z_m(0) \log \frac{e}{1} + z_m(0) \log \frac{e}{e} \right) \\ = & h \sum_{k=1}^{D} z_k(0) - h x_m(0) \\ = & h(1 - x_m(0)). \quad \square \end{aligned}$$

□

Proof of (15)**,** (18) **and** (19) To derive equation (15) we only have to replace the simplex operators in the equation $\boldsymbol{x}(h) = \boldsymbol{x}(0) \oplus h \odot \mathbf{u}$ by their definition. Since we use ratios of the components of the same CoDa vector on each side of the equation we can omit the closure, allowing us to derive

$$\frac{x_i(h)}{x_j(h)} = \frac{x_i(0)}{x_j(0)} (\frac{u_i}{u_j})^h = \frac{x_i(0)}{x_j(0)} \exp(h \log \frac{u_i}{u_j}). \quad \square \tag{45}$$

The results of (18) and (19) correspond to (45) for the direction $\boldsymbol{u} = \boldsymbol{e_m}$. Since (45) only involves rations of the elements of u we may equivalently the non-closed direction $\boldsymbol{u} = (1, \ldots, e, \ldots, 1)$, leading to

$$\begin{aligned} \text{if} \qquad l \neq m, \frac{x_m(h)}{x_l(h)} &= \frac{x_m(0)}{x_l(0)} \exp(h \log \frac{e}{1}) = \frac{x_m(0)}{x_l(0)} \exp(h) \quad \square \\ \text{if} \qquad j, l \neq m, \frac{x_j(h)}{x_l(h)} &= \frac{x_m(0)}{x_l(0)} \exp(h \log \frac{1}{1}) = \frac{x_j(0)}{x_l(0)} \exp(h). \quad \square \end{aligned}$$

□

Proof of (23) Let us first note that we can rewrite (11) as

$$\mathbf{X_m}(h) = \frac{X_m(0)}{\sum_{k=0}^{D_X} X_k(0) \exp h \log(u_k).}$$

Taking the log and then the derivative, we get

$$\frac{\partial \log X_m(h)}{\partial h} = \log u_m - C(h), \tag{46}$$

where $C(h) = \frac{\sum_{k=0}^{D_X} X_k(0) \log u_k \exp(h \log(u_k))}{\sum_{k=0}^{D_X} X_k(0) \exp(h \log(u_k)}$ is independent from m.

A Taylor expansion of $f_j(h) := \underline{f}_j(\boldsymbol{x}(0) \oplus h \odot \boldsymbol{u})$ at $h = 0$ for all $j = 1, \dots D_Y$ yields

$$\begin{aligned} f_j(h) &\simeq f_j(0) + h \sum_{m=0}^{D_X} \frac{\partial f_j}{\partial \log f_j} \frac{\partial \log f_j}{\partial \log \check{x}_m} |_{h=0} \frac{\partial \log \check{x}_m(h)}{\partial h} |_{h=0} \\ &\simeq f_j(0) + h \sum_{m=0}^{D_X} f_j(0) \frac{\partial \log \tilde{Y}_j}{\partial \log \check{x}_m} |_{h=0} [\log u_m - C(0)] \end{aligned}$$

Using (46) we see that the term involving $C(0)$ cancels out when taking the closure leading to (23). □

Proof of (29) From (46) in the proof of (23), we easily derive that

$$\log X_{j'}(h) - \log X_{l'}(h) = h \log \frac{u_{j'}}{u_{l'}}.$$

□

Proof of (30) From (20), it clear that $\frac{dY}{dh} =< \operatorname{clr} \boldsymbol{\beta}, \operatorname{clr} \boldsymbol{u} >_A$. It is then enough to combine with (29).

Proof of (31) Another expression can be obtained for the left hand side of (30). Indeed

$$\begin{aligned} \frac{\frac{dY}{dh}}{\frac{d \log \frac{X_j}{X_l}}{dh}} &= \frac{\frac{dY}{dh}}{\frac{d \log X_j}{dh}} \frac{d \log X_j}{d \log \frac{X_j}{X_l}} + \frac{\frac{dY}{dh}}{\frac{d \log X_l}{dh}} \frac{d \log X_l}{d \log \frac{X_j}{X_l}} \\ &= \frac{SE_j \log(u_j) + SE_l \log(u_l)}{\log(u_j) - \log(u_l)} \end{aligned}$$

It is then enough to apply this result to the vector $\mathbf{e}_j \ominus \mathbf{e}_l$ and we thus obtain the difference $SE_j - SE_l$.

Proof of (33) The first expression is obtained by taking the log and then the derivative of (32) and the second expression is a consequence of writing the log of the ratio as a difference of logs.

Acknowledgments

We acknowledge funding from the French National Research Agency (ANR) under the Investments for the Future (Investissements d'Avenir) program, grant ANR-17-EURE-0010, from the French National Association of Research and Technology (ANRT) and from the market research agency BVA.

References

[Aitchison, 1986] Aitchison, J. (1986). The statistical analysis of compositional data. Chapman and Hall, London.

[Aitchison and Shen, 1980] Aitchison, J. and Shen, S. M. (1980). Logistic-normal distributions: Some properties and uses. Biometrika, 67(2):261–272.

[Boogaart and Tolosana-Delgado, 2013] Boogaart, K. G. v. d. and Tolosana-Delgado, R. (2013). Analyzing Compositional Data with R. Springer Science & Business Media.

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

[Coenders et al., 2017] 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(6):494–512.

[Coenders and Pawlowsky-Glahn, 2020] Coenders, G. and Pawlowsky-Glahn, V. (2020). On interpretations of tests and effect sizes in regression models with a compositional predictor. SORT, 44(1).

[Dargel and Thomas-Agnan, 2023] Dargel, L. and Thomas-Agnan, C. (2023). The link between multiplicative competitive interaction models and compositional data regression with a total. TSE Working Paper N°1455.

[Daunis-i Estadella et al., 2002] Daunis-i Estadella, J., Egozcue, J. J., and Pawlowsky-Glahn, V. (2002). Least squares regression in the simplex. In Proceedings of IAMG, pages 411–416.

[Dumuid et al., 2019] 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(3):846–857.

[Egozcue et al., 2012] Egozcue, J. J., Daunis-I-Estadella, J., Pawlowsky-Glahn, V., Hron, K., and Filzmoser, P. (2012). Simplicial regression. the normal model. Journal of Applied Probability and Statistics, 6(1-2):87–108.

[Ferrer-Rosell et al., 2016] Ferrer-Rosell, B., Coenders, G., Mateu-Figueras, G., and Pawlowsky-Glahn, V. (2016). Understanding low-cost airline users' expenditure patterns and volume. Tourism Economics, 22(2):269–291.

[Hron et al., 2012] Hron, K., Filzmoser, P., and Thompson, K. (2012). Linear regression with compositional explanatory variables. Journal of Applied statistics, 39(5):1115–1128.

[Jašková et al., 2023] Jašková, P., Palarea-Albaladejo, J., Gába, A., Dumuid, D., Pedišić, Ž., Pelclová, J., and Hron, K. (2023). Compositional functional regression and isotemporal substitution analysis: Methods and application in time-use epidemiology. Statistical Methods in Medical Research, page 09622802231192949.

[Katz and King, 1999] Katz, J. N. and King, G. (1999). A statistical model for multiparty electoral data. American Political Science Review, 93(1):15–32.

[Morais and Thomas-Agnan, 2021] Morais, J. and Thomas-Agnan, C. (2021). Impact of covariates in compositional models and simplicial derivatives. Austrian Journal of Statistics, 50(2):1–15.

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

[Müller et al., 2016] Müller, I., Hron, K., Fišerová, E., Šmahaj, J., Cakirpaloglu, P., and Vancakova, J. (2016). Time budget analysis using logratio methods. arXiv preprint arXiv:1609.07887.

[Nguyen et al., 2022] Nguyen, T. H. A., Laurent, T., Thomas-Agnan, C., and Ruiz-Gazen, A. (2022). Analyzing the impacts of socio-economic factors on french departmental elections with coda methods. Journal of Applied Statistics, 49(5):1235–1251.

[Nguyen et al., 2020] Nguyen, T. H. A., Thomas-Agnan, C., Laurent, T., and Ruiz-Gazen, A. (2020). A simultaneous spatial autoregressive model for compositional data. Spatial Economic Analysis.

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

[Ruiz-Gazen et al., 2023] Ruiz-Gazen, A., Thomas-Agnan, C., Laurent, T., and Mondon, C. (2023). Detecting outliers in compositional data using invariant coordinate selection check for updates. Robust and Multivariate Statistical Methods: Festschrift in Honor of David E. Tyler, page 197.

[Thomas-Agnan et al., 2023] Thomas-Agnan, C., Laurent, T., and Ruiz-Gazen, A. (2023). Covariates impacts in spatial autoregressive models for compositional data. to appear in Journal of Spatial Econometrics.

[Trinh et al., 2018] Trinh, H. T., Morais, J., Thomas-Agnan, C., and Simioni, M. (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, page 0962280218770223.

[van den Boogaart et al., 2021] van den Boogaart, K., Filzmoser, P., Hron, K., Templ, M., and Tolosana-Delgado, R. (2021). Classical and robust regression analysis with compositional data. Mathematical geosciences, 53:823–858.

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