STRUCTURAL ZEROS IN REGRESSION MODELS WITH COMPOSITIONAL EXPLANATORY VARIABLES

Francesco Porro [1]

[1] Dipartimento di Matematica, Università degli Studi di Genova,
e-mail: francesco.porro@unige.it

ABSTRACT: In many real-life situations it may happen to consider a regression model with compositional explanatory variables. Compositional data describe parts of some whole, having the feature to sum to a fixed value, so they are commonly presented as vectors of proportions, percentages, or frequencies. In the compositional framework, the presence of structural zeros in the regressors is problematic, since a composition is not allowed to have a part equal to zero. In the recent years, a few techniques have been introduced in the literature to adress this issue. In this paper a description and a comparison of the most interesting proposals are provided.

KEYWORDS: Compositional data, regression models, structural zeros, logratio transformation.

1 The compositional data framework

During the last decades Compositional Data (CoDa) have gained more attention in the literature. The relevant information in compositional data is in the ratios between the parts and not in their absolute values or in their sum. Different examples of compositional data can be easily found in every field: physics, chemistry, finance, social sciences, and economics, just to mention some of them (cf. Pawlowsky-Glahn *et al.* , 2015). The CoDa methodology has been developed to deal with the compositions.

Definition 1 *Let $D \in \mathbb{N}$. Consider the real-valued vectors $\mathbb{R}^D$, with all (strictly) positive components. Two of such vectors $\mathbf{x} = (x_1, x_2, \ldots, x_D)$ and $\mathbf{y} = (y_1, y_2, \ldots, y_D)$ are compositionally equivalent whether there exists a positive constant $c \in \mathbb{R}$ such that $\mathbf{x} = c \cdot \mathbf{y}$. A D-part composition is then a class of equivalence containing all the compositionally equivalent vectors in $\mathbb{R}^D$.*

Since a D-part composition is an equivalence class, a representative one has to be selected: it is usually the vector of proportions that sum to 1. The sample

space for the D-part compositions is the simplex $\mathbb{S}^D$, defined by:

$$\mathbb{S}^D = \{(x_1, x_2, \ldots, x_D) \in \mathbb{R}^D \ : \ x_i > 0 \ \forall i; \ \sum_{i=1}^{D} x_i = c\}, \tag{1}$$

where the arbitrary constant c is usually set to 1. For further details, see Pawlowsky-Glahn *et al.* , 2015, Filzmoser *et al.* , 2018, and references therein. Starting from the definition of composition, the so-called *Aitchison geometry on the simplex* can be defined: it is the suited framework to analyze compositional data, and it can be equipped with a coherent distance, norm, and inner product. In CoDa analysis, a dataset $\mathbf{X}$ is a sample of n observations, each one being a D-part composition $\mathbf{X} = (\mathbf{x}_1, \mathbf{x}_2, \ldots, \mathbf{x}_n)'$, with $\mathbf{x}_i = (x_{i1}, x_{i2}, \ldots, x_{iD})$, $i = 1, 2, \ldots, n$. The standard statistical descriptive measures, based on the real Euclidean structure, should be used with attention in such a dataset, since they can lead to erroneous conclusions (see Pawlowsky-Glahn *et al.* , 2015). To overcome this issue, the compositional approach proposes alternative statistical tools and methods, based on the Aitchison geometry.

A usual practice in handling compositions is the application of transformations, mapping them into real vectors (belonging to suitable spaces) for exploiting the usual Euclidean structure. Several transformations based on logratios have been proposed: the additive (*alr*), the centered (*clr*) and the isometric (*ilr*) logratio transformations. Their features can be found in Pawlowsky-Glahn *et al.* , 2015 and Filzmoser *et al.* , 2018. The definition of *ilr*-transformation is the following one.

Definition 2 *For a D-part composition $\mathbf{x} = (x_1, x_2, \ldots, x_D)$, the isometric logratio (ilr) transformation associated to an Aitchison-orthonormal basis in $\mathbb{S}^D$, $\{\mathbf{e}_i\}$, $(i = 1, 2, ..., D-1)$, is the mapping from $\mathbb{S}^D$ to $\mathbb{R}^{D-1}$ given by:*

$$ilr(\mathbf{x}) = [\langle \mathbf{x}, \mathbf{e}_1 \rangle_a, \langle \mathbf{x}, \mathbf{e}_2 \rangle_a, ..., \langle \mathbf{x}, \mathbf{e}_{D-1} \rangle_a],$$

where $\langle \cdot, \cdot \rangle_a$ denotes the Aitchison inner product in $\mathbb{S}^D$, defined by:

$$\langle \mathbf{x}, \mathbf{y} \rangle_a = \frac{1}{2D} \sum_{i=1}^{D} \sum_{j=1}^{D} \left(\ln \frac{x_i}{x_j} \ln \frac{y_i}{y_j} \right).$$

For the remainder of this paper, it is worth just mentioning that the *ilr*-transformation is characterized by two important features: (i) it reduces the number of parts, since a D-part composition is mapped into a vector in $\mathbb{R}^{D-1}$; (ii) it preserves both the distances and the angles: in the simplex the Aitchison distance of two compositions is equal to the distance of the corresponding *ilr*-transformed vectors in $\mathbb{R}^{D-1}$ (see Pawlowsky-Glahn *et al.* , 2015 for details).

2 Regression models with compositional regressors

Many examples of regression models with (at least some) compositional explanatory variables can be easily found. In such a case, the regressors can not be directly used since compositional data are by definition singular: the constraint about their sum provides the linear dependency of the regressors and a singular covariance matrix.

The standard approach is to apply the *ilr*-transformation to the original explanatory variables and to consider the corresponding *ilr*-transformed variables as new regressors (Hron *et al.* , 2012). In this way, the linear dependence of the compositional regressors can be discarded: the new obtained model can be easily handled, and then parameter estimation can be done as in usual linear regression. This approach cannot be applied whether there are zeros, since in this case, no logratio transformation can be carried out. It follows that in case of *structural zeros* in the regressors, a different procedure has to be undertaken. It is worth recalling that a structural zero is a value that is certain to be zero, and it is not due to imprecise or insufficient measurements.

3 Three approaches dealing with structural zeros

For facing the issue of structural zeros in regression models with compositional regressors a few approaches have been proposed, quite recently. In the following, three of them are briefly presented: the first one is due to Aitchison, 1986, while the other two are described in Verbelen *et al.* , 2018. In the presentation more details will be provided to characterize and compare the three methods.

3.1 A naive approach: the replacement

The replacement strategy is the first method proposed in the literature, and it is the most intuitive one. The idea is to take all those values giving problems (since, for example, they are zeros) and replace them with a nonproblematic value (for example, a value very close to zero). This approach can be very easily implemented, and it can also be used to remove missing values. The most relevant drawbacks are the arbitrary nature of the replaced values, and the inconsistency in case of structural zeros, as they are *true* zeros.

3.2 The conditioning approach

The conditioning approach consists in treating the observations with different structural 0 patterns as different subgroups within the data, so that the regression coefficients are modeled conditionally on the 0 patterns. This method requires to compute for each compositional observation with at least one 0 part, the *ilr*-transformation of the corresponding subcomposition with non-zero parts (obtained by removing the zero parts) and to model the compositional predictor effect separately by 0 pattern. The regression coefficients obtained by this method are different for each structural 0 pattern and hence only estimated by using observations with that particular 0 pattern.

3.3 The projection approach

The projection approach is more parsimonious than the conditional one, since the regression parameters are shared across the different 0 patterns. In this method, a *generalized isometric logratio transformation* from the simplex $\mathbb{S}^D$ to $\mathbb{R}^{D-1}$ is proposed as an extension of the usual one. This new transformation can be applied also to a compositions with one or more zero parts, since the logratios are calculated using the projections onto the orthogonal complement of the structural 0 parts.

References

AITCHISON, JOHN. 1986. *The Statistical Analysis of Compositional Data*. Chapman and Hall.

FILZMOSER, PETER, HRON, KAREL, & TEMPL, MATTHIAS. 2018. *Applied compositional data analysis*. Springer.

HRON, KAREL, FILZMOSER, PETER, & THOMPSON, K. 2012. Linear regression with compositional explanatory variables. *J.Appl.Stat.*, **39**(5), 1115–1128.

PAWLOWSKY-GLAHN, VERA, EGOZCUE, JUAN JOSÉ, & TOLOSANA-DELGADO, RAIMON. 2015. *Modeling and analysis of compositional data*. John Wiley & Sons.

VERBELEN, ROEL, ANTONIO, KATRIEN, & CLAESKENS, GERDA. 2018. Unravelling the predictive power of telematics data in car insurance pricing. *J. R. Stat. Soc., C: Appl. Stat.*, **67**(5), 1275–1304.