Aggregate data and compositional variables

Enora Belz[a], Arthur Charpentier[b]
Faculté des Sciences Économiques, Université de Rennes 1
[a]enora.belz@univ-rennes1.fr, [b]arthur.charpentier@univ-rennes1.fr

Introduction

In many applications, **individual** data are not widely available, if at all, especially when small geographic areas or vulnerable data are involved. It is more frequent to find **aggregated** data instead. Two problems arise :

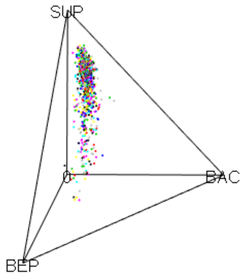
- Can these aggregated data be used to **infer** individual behavior (we will use the term of **ecological inference**) ?
- How to **manipulate** categorical variables (which become **compositional variables** once aggregated) ?

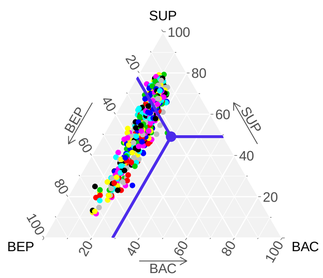
Compositional Data

A **composition** of D components is a vector $\mathbf{x}$ of the simplex S^D defined as

$$S^D = \left\{ \mathbf{x} = [x_1, x_2, \ldots, x_D];\ x_i > 0, i = 1, 2, \ldots, D;\ \sum_{i=1}^{D} x_i = 1 \right\} \quad (1)$$

This sample space is a $(D-1)$-dimensional subset of $\mathbb{R}^D$. For example, S^3 is a triangle (**ternary diagram**).





Picpus (12^e) : 22% BEP, 29% BAC, 79% SUP

Aitchison (1986) [1] introduces operations on the simplex (sum and scalar multiplication). The **inner product** of two compositions is defined as :

$$\langle \mathbf{x}, \mathbf{y} \rangle_a = \sum_{i=1}^{D} \log \frac{x_i}{g(\mathbf{x})} \log \frac{y_i}{g(\mathbf{y})} = \frac{1}{D} \sum_{i<j} \log \frac{x_i}{x_j} \log \frac{y_i}{y_j} \quad (2)$$

The **additive** log-ratio transformation :

$$alr(\mathbf{x}) = [\log \frac{x_1}{x_D}, \log \frac{x_2}{x_D}, \ldots, \log \frac{x_{D-1}}{x_D}] \quad (3)$$

The **centered** log-ratio transformation :

$$clr(\mathbf{x}) = [\log \frac{x_1}{g(\mathbf{x})}, \log \frac{x_2}{g(\mathbf{x})}, \ldots, \log \frac{x_D}{g(\mathbf{x})}] \quad (4)$$

The **isometric** log-ratio transformation :

$$ilr(\mathbf{x}) = [\langle \mathbf{x}, e_1 \rangle_a, \langle \mathbf{x}, e_2 \rangle_a, \ldots, \langle \mathbf{x}, e_{D-1} \rangle_a] \quad (5)$$

Where $e_1, e_2, ..., e_{D-1}$ represent an orthonormal basis of S^D and $g(\mathbf{x}) = (x_1 x_2 \ldots x_D)^{1/D}$ is the geometric mean.

Econometrics with Compositional Data

Regression analysis is generally used in statistics to study the relation between a response Y and a set of explanatory covariate $\mathbf{x} = (x_1, \ldots, x_D)$ as

$$\mathbf{y} = \beta_0 + \beta^T \mathbf{x} + \epsilon \quad (6)$$

The standard linear model is **unreasonable** with compositional data.

- Compositions contain only **relative information** : least squares methods examine continuous variables that are linked with an absolute relationship
- When one part is altered, **another is altered** : the interpreation of the linear regressions coefficients assumes that "all other things being equal"
- Composition covariates involves **multicollinearity** problems
- The simplex has **particular geometric** properties and operations : most common statistical procedures are developed in the usual Euclidean geometry

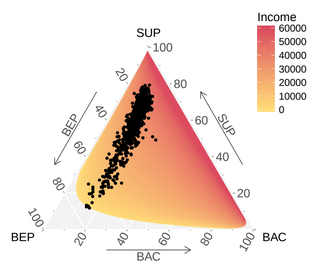
The ilr transformation is an isometric transformation from the simplex to the real space with usual Euclidean geometry. It leads to allow the application of the standard models to the compositions as

$$\mathbf{y} = \beta_0 + \beta^T ilr(\mathbf{x}) + \epsilon \quad (7)$$

Using the estimation of β is possible to find b with the ilr-inverse transformation. The composition b can be interpreted as the slope parameter of the standard regression. If x differs by $\frac{b}{\|b\|}$ in the direction of b, then y differs by $\|b\|$.

An example

Analysis of median income according to diploma levels in Paris by iris.



Median income increases in the SUP direction and decreases in the BEP direction. In other words, when the proportion of SUP increases or the proportion of BEP decreases in an area, then median income increases.

Ecological inference

Ecological inference consists of making inference of the **individual** behaviour using **aggregate** group-level data. Group-level data are usually less reliable and subject to bias and imprecision that may involve **mistake of inference**. Robinson (1950) [5] pointed out to be careful with the use of aggregate data to study individuals (**"ecological fallacy"**).

	Republican	Democrate	Total
Men	?	?	ϕ_j
Women	?	?	$1-\phi_j$
Total	p_j	$1-p_j$	1

The aim is to find β_j^0 and β_j^1 which are the probability of being republican conditionally of being a man or a woman in the area j.
Goodman regression (1953) [3] suggests that these two probabilities are **constant** over area and could be estimated by least squares as

$$p_j = \beta^1 \phi_j + \beta^0 (1 - \phi_j) \quad (8)$$

The method of bound (Duncan and Davis (1953) [2]) is to find the minimum and maximum of the probability with

$$\max\left\{0; \frac{p_j - (1-\phi_j)}{\phi_j}\right\} \leq \beta_j^1 \leq \min\left\{\frac{p_j}{\phi_j}; 1\right\} \quad (9)$$

$$\max\left\{0; \frac{p_j - \phi_j}{1-\phi_j}\right\} \leq \beta_j^0 \leq \min\left\{\frac{p_j}{1-\phi_j}; 1\right\} \quad (10)$$

King's solution (1997) [4] is an improvement in ecological inference by combining the Goodman method and the information of the bounds to improve inference. β_j^0 and β_j^1 are linked by **tomography line** within the unit square as

$$\beta_j^0 = \frac{p_j}{1-\phi_j} - \frac{\phi_j}{1-\phi_j} \beta_j^1 \quad (11)$$

King suggests three assumptions :

- β_j^0 and β_j^1 are in a single cluster which is generated by a **truncated normal** bivariate distribution conditional of ϕ_j
- Absence of **spatial autocorrelation** : the number of exposed cases of an area not depends of the number exposed cases in the others area
- Absence of **aggregation bias** : independence between the regressors ϕ_j and the parameters β_j^0 and β_j^1

References

[1] J. Aitchison. *The Statistical Analysis of Compositional Data*. Chapman and Hall London, 1986.
[2] O. Duncan and B. Davis. An Alternative to Ecological Correlation. *American Sociological Review*, 18(6):665–666, dec 1953.
[3] L. Goodman. Ecological Regressions and Behavior of Individuals. *American Sociological Review*, 18(6):663–664, dec 1953.
[4] G. King. *A Solution to the Ecological Inference Problem*. Princeton University Press, 1997.
[5] W. Robinson. Ecological Correlation and the Behavior of Individuals. *American Sociological Review*, 15(3):351–357, jun 1950.