CODA methods and the multivariate Student distribution: an application to political economy

Thi Huong An Nguyen [12] & Thibault Laurent [1]

[1] *Toulouse School of Economics, (huongan.nguyen, thibault.laurent)@tse-fr.eu*
[2] *Danang University of Architecture, Vietnam, annth@dau.edu.vn*

Résumé. Le vecteur des proportions de vote par parti sur une subdivision donnée d'un territoire est un vecteur de données dites de composition (mathématiquement, un vecteur appartenant à un simplexe). Les économistes politiques s'intéressent à l'impact des caractéristiques socio-économiques des unités géographiques sur le résultat des élections. Parce que les données de parts de votes présentent souvent davantage de valeurs extrêmes que des observations issues d'une loi normale, nous avons décidé d'utiliser une distribution d'erreur de Student dans le modèle de régression. Nous décrivons comment adapter le modèle de régression CODA à la distribution d'erreur multivariée de Student. Pour un vecteur d'erreur gaussien, l'hypothèse de coordonnées indépendantes équivaut celle de coordonnées non corrélées. Cependant, cette équivalence n'est plus vraie lorsqu'on envisage une distribution multivariée de Student. Dans cet article, nous nous concentrons sur la construction d'un modèle de régression CODA ayant des vecteurs d'erreurs de loi de Student multivariées indépendantes. Les modèles sont ajustés aux données électorales françaises des élections départementales de 2015.

Mots-clés. Distribution de Student multivariée indépendante, distribution de Student multivariée non corrélée, modèles de rgression compositionnelle, estimateur de maximum vraisemblance, queue lourde, ...

Abstract. In a multiparty election, the vote shares form a composition vector (mathematically, a vector belonging to a simplex). Political economists are interested by the impact of the socio-economic characteristics of the geographical units on the outcome of the elections. Because vote shares data often exhibit heavy tail behavior, we decide to use a Student error distribution. We describe how to adapt the CODA regression model to the multivariate Student error distribution. For a Gaussian errors vector, the assumption of independent coordinates is equivalent to the assumption of uncorrelated coordinates. However, this equivalence is no longer true when considering a multivariate Student distribution. In this paper, we concentrate on building a CODA regression model with multivariate independent Student error vectors. The models are fitted on French electoral data of the 2015 departmental elections.

Keywords. Independent multivariate Student distribution, Uncorrelated multivariate Student distribution, compositional regression models, Maximum Likelihood Estimator, heavy tail, ...

1 Introduction

Recently, many authors in political economy concentrate on building models and understanding the drivers of the outcome of a two-party electoral system (Beauguitte and Colange (2013), Ansolabehere and Leblanc (2008)). The outcome of an election can be influenced by the campaign strategies of candidates, demographic factors such as age, domain of activity, rate of unemployment, and so on. In this work, we are interested in exploring the impact of the characteristics of the demographics and social factors on the outcome of the 2015 French departmental election. The outcomes of the election in this multiparty system consist of vectors whose components are the proportions of votes per party. In what follows, our attention focuses on the relationship between votes shares and socio-economics factors such as age, education levels, domain of activities, unemployment rate and so on by using CODA (COmpositional Data Analysis) regression models.

In the statistical literature, there are regression models adapted to share vectors including CODA models, but also Dirichlet models, Student models and others. In these models, the dependent and independent variables may be compositional variables (see Mert et al (2016)). Honaker et al (2002), Katz and King (1999) use a statistical model for multiparty electoral data assuming that the territorial units yield independent observations. Morais et al (2017) study the impact of media investments on brand's market shares with a CODA regression model. Nguyen et al (2018) study a CODA multivariate regression model which uses the normal distribution to illustrate the impacts of socio-economic factors on French departmental elections. However, this election data often exhibit heavy tail behavior (see Katz and King (1999)). In order to eliminate the heavy tail problem, a proposal found in the literature is to replace the Gaussian distribution by the Student distribution.

In one dimension, the generalized Student distribution allows for heavier tails when the shape parameter is small. In higher dimensions, there are several kinds of multivariate Student models (see Johnson and Kotz (1972) and Kotz for overview). There are two versions of the multivariate Student distribution: the independent Student (IT) and the uncorrelated Student (UT) (see Prucha and Kelejian (1985)). Nguyen et al (2019) present a full summary of these two versions. They consider a multivariate dependent vector and a linear regression model with three different assumptions on the error term distribution: the Gaussian distribution ($\boldsymbol{\epsilon}_N$), The Uncorrelated Student distribution ($\boldsymbol{\epsilon}_{UT}$), the Independent Student distribution ($\boldsymbol{\epsilon}_{IT}$). Nguyen et al (2019) derive some theoretical properties of the UT model and propose a simple iterative reweighted algorithm to compute the maximum likelihood estimators in the IT model. However, Nguyen et al (2019) show that the UT model has limitation of assumption of the single realization. This restricts the properties of the maximum likelihood estimators and prevent the use of tests against the other two models. Thus, we will concentrate in multivariate IT case in this paper.

Vote share data of the 2015 French departmental election for 95 departments in France

Table 1: Data description.

Variable name	Description	Averages
Vote share	Left(L), Right(R), Extreme Right(XR)	0.37, 0.388, 0.242
Age	Age_1840, Age_4064, Age_65.	0.313, 0.432, 0.255
Diploma	<BAC, BAC, SUP.	0.591, 0.16, 0.239
Employment	AZ, BE, FZ, GU, OQ	0.031, 0.099, 0.049, 0.439, 0.382
unemp	The unemployment rate	0.117
employ_evol	Mean annual growth rate of employment (2009-2014)	-0.145
owner	The proportion of people who own assets	0.616
income_ tax	The proportion of people who pay income tax	0.552
foreign	The proportion of foreigners	0.050

are collected from the Cartelec website [1] and corresponding socio-economic data (for 2014) have been downloaded from the INSEE website [2]. Table 1 summarizes our data set and see Nguyen et al (2018).

2 Compositional regression models

2.1 Principles of compositional data analysis

A composition $\mathbf{x}$ is a vector of D parts of some whole which carries relative information. A D-composition $\mathbf{x}$ lies in the so-called simplex space $\mathbf{S}^D$ defined by:

$$\mathbf{S}^D = \{\mathbf{x} = (x_1, ..., x_D)' : x_j > 0, j = 1, ..., D; \sum_{j=1}^{D} x_j = 1\}$$

The vector space structure of the simplex $\mathbf{S}^D$ is defined by the perturbation and powering operations:

$$\mathbf{x} \oplus \mathbf{y} = \mathcal{C}(x_1 y_1, \ldots, x_D y_D),\ \mathbf{x},\ \mathbf{y} \in \mathbf{S}^D$$
$$\lambda \odot \mathbf{x} = \mathcal{C}(x_1^\lambda, \ldots, x_D^\lambda),\ \lambda \text{ is a scalar}, \mathbf{x} \in \mathbf{S}^D.$$

where $\mathcal{C}(\mathbf{x}) = \left(\frac{x_1}{\sum_{j=1}^D x_j}, \cdots, \frac{x_D}{\sum_{j=1}^D x_j} \right)$ is the closure operation.

The compositional matrix product, corresponding to the matrix product in the simplex, is defined by

$$\mathbf{B} \boxdot \mathbf{x} = \mathcal{C}\left(\prod_{j=1}^{D} x_j^{b_{1j}}, \cdots, \prod_{j=1}^{D} x_j^{b_{Lj}} \right)^T$$

[1] https://www.data.gouv.fr/fr/datasets/elections-departementales-2015-resultats-par-bureaux-de-vote/

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

where $\mathbf{B} = (b_{lj}),\ l = 1, \ldots, L,\ j = 1, \ldots, D$, is a parameter matrix such that the column vectors belong to $\mathbf{S}^D, \mathbf{j}_L^T\mathbf{B} = \mathbf{0}_D,\ \mathbf{B}\mathbf{j}_D = \mathbf{0}_L$, where $\mathbf{j}_L$ is a $L \times 1$ column vector of ones, and $\mathbf{j}_L^T$ is the transposed of $\mathbf{j}_L$.
The simplex $\mathbf{S}^D$ can be equipped with the Aitchison inner product (Aitchison (1985) and Pawlowsky-glahn (2015)) in order to define distances. The compositional inner product $\langle \mathbf{x}, \mathbf{y} \rangle_c$ and the expected value $\mathbb{E}^{\oplus}\mathbf{Y}$ are also defined in Pawlowsky-glahn (2015).

To define our regression model, we choose to work with ilr transformations from the simplex space $\mathbf{S}^D$ to the Euclidean space $\mathbb{R}^{D-1}$. An isometric log-ratio transformation (ilr) is defined by $\text{ilr}(\mathbf{x}) = \mathbf{V}_D^T \ln(\mathbf{x})$ where the logarithm of $\mathbf{x}$ is understood componentwise, $\mathbf{V}_D^T$ is a transposed contrast matrix (Pawlowsky-glahn (2015)) associated to a given orthonormal basis $(\mathbf{e}_1, \cdots, \mathbf{e}_{D-1})$ of $\mathbf{S}^D$ by $\mathbf{V}_D = \text{clr}(\mathbf{e}_1, \cdots, \mathbf{e}_{D-1})$. As in Pawlowsky-glahn (2015), we use the following contrast matrix for $D = 3$

$$\mathbf{V}_3 = \begin{bmatrix} 2/\sqrt{6} & 0 \\ -1/\sqrt{6} & 1/\sqrt{2} \\ -1/\sqrt{6} & -1/\sqrt{2} \end{bmatrix}$$

to define the log-ratio transformations. This particular matrix defines the following ilr coordinates

$$\text{ilr}_1(\mathbf{x}) = \frac{1}{\sqrt{6}}(2\log x_1 - \ln x_2 - \log x_3) = \frac{2}{\sqrt{6}} \log \frac{x_1}{\sqrt{x_2 x_3}}$$
$$\text{ilr}_2(\mathbf{x}) = \frac{1}{\sqrt{2}}(\log x_2 - \log x_3) = \frac{1}{\sqrt{2}} \log \frac{x_2}{x_3}$$

The first ilr coordinate contains information about the relative importance of the first component x_1 with respect to the geometric mean of the second and the third components $g = \sqrt{x_2 x_3}$. The second ilr coordinate contains information about the relative importance of the second component x_2 with respect to the third component x_3. In our case, the first ilr coordinate opposes the Left wing to the group of the Right wing and the Extreme Right party and the second opposes the Right wing to the Extreme Right party.

2.2 CODA regression models

In this paper, we use the notations in Table 2. Let $\mathbf{Y}_i$ denotes the compositional response value of the ith observation, $\mathbf{Y}_i \in \mathbf{S}^L$, and $\mathbf{X}_i^{(q)},\ q = 1, \ldots, Q$, denotes the value of the qth compositional covariate for the ith observation, $\mathbf{X}_i^{(q)} \in \mathbf{S}^{D_q},\ q = 1, \ldots, Q$, $Z_{ki},\ k = 1, \ldots, K$, denotes the kth classical covariate of the ith observation. Let us first introduce the CODA regression model in the ilr coordinate space as follows:

$$\text{ilr}(\mathbf{Y}_i) = \mathbf{b_0}^* + \sum_{q=1}^{Q} \text{ilr}(\mathbf{X}_i^{(q)})\mathbf{B}_q^* + \sum_{k=1}^{K} Z_{ki}\mathbf{c}_k^* + \text{ilr}(\boldsymbol{\epsilon}_i) \tag{1}$$

Table 2: Notations

Variable	Notation	Coordinates
Dependent	$\mathbf{Y}_i = (Y_{i1}, \ldots, Y_{iL})$	$\text{ilr}(\mathbf{Y}_i) = \mathbf{Y}_i^*$
Compositional explanatory	$\mathbf{X}_i^{(q)} = (X_{i1}^{(q)}, \ldots, X_{iD_q}^{(q)})$	$\text{ilr}(\mathbf{X}_{ip}^{(q)}) = \mathbf{X}_{ip}^{(q)*}$
Classical explanatory	Z_{ki}	
General notations		
L	Number of components of the dependent variable	
$i = 1, \ldots, n$	Index of observations ($n = 95$)	
$q = 1, \ldots, Q$	Index of compositional explanatory variables ($Q = 3$)	
$p = 1, \ldots, D_q$	Index of the coordinates for the compositional explanatory variables	
$k = 1, \ldots, K$	Index of classical explanatory variables ($K = 5$)	

where $\text{ilr}(\mathbf{Y_i})$, $\text{ilr}(\mathbf{X}_i^{(q)})$ are the ilr coordinates of $\mathbf{Y}_i$, $\mathbf{X}_i^{(q)}$ $(q = 1, \ldots, Q)$ respectively, $\text{ilr}(\mathbf{Y_i}) \in \mathbb{R}^{L-1}$, $\text{ilr}(\mathbf{X}_i^{(q)}) \in \mathbb{R}^{D_q - 1}$; $\mathbf{b_0}^*$, $\mathbf{B}_q^*$, $\mathbf{c}_k^*$ are the parameters in the coordinate space, and $\text{ilr}(\boldsymbol{\epsilon}_i)$ are the residuals in the coordinate space, $\text{ilr}(\boldsymbol{\epsilon}_i) \in \mathbb{R}^{L-1}$. The classical distributional assumption is that $\text{ilr}(\boldsymbol{\epsilon})$ follows a multivariate Gaussian distribution. However our case, we will assume that $\text{ilr}(\boldsymbol{\epsilon})$ follows an independent multivariate Student (IT) distribution with zero mean and covariance matrix $\boldsymbol{\Sigma}_{IT}$.
Let $\bigoplus$ denotes the summation, this regression model (1) can be written in the simplex as

$$\mathbf{Y}_i = \mathbf{b_0} \bigoplus_{q=1}^{Q} \mathbf{B}_q \boxdot \mathbf{X}_i^{(q)} \bigoplus_{k=1}^{K} Z_{ki} \odot \mathbf{c}_k \oplus \boldsymbol{\epsilon}_i, \quad i = 1, \ldots, n \tag{2}$$

where $\mathbf{b_0}, \mathbf{B}_1, \ldots, \mathbf{B}_Q, \mathbf{c}_1, \ldots, \mathbf{c}_K$ are the parameters satisfying $\mathbf{b_0} \in \mathbf{S}^L$, $\mathbf{B}_q \in \mathbf{S}^{D_q}$, $q = 1, \ldots, Q$, $\mathbf{c}_k \in \mathbf{S}^L$, $k = 1, \ldots, K$, $\mathbf{j}_L^T \mathbf{B}_q = \mathbf{0}_{D_q}$, $\mathbf{B}_q \mathbf{j}_{D_q} = \mathbf{0}_L$. The distributional assumption is that $\boldsymbol{\epsilon}_i \in \mathbf{S}^L$ follows the independent multivariate Student (IT) distribution on the simplex. We estimate the parameters of this model using the iterative reweighting algorithm described in Nguyen et al (2019). Table 3 shows that the multivariate Student regression model explains better than the multivariate Gaussian regression model. Moreover, a test based on the Mahalanobis distance shows that we do not reject the null hypothesis of the Student distribution, but we do reject the null hypothessis of the Gaussian one.

Table 3: Multivariate Gaussian and Student regression models with compositional and classical variables

	Gaussian model		*Student model, $\nu = 4$*	
	y_ilr[, 1]	y_ilr[, 2]	y_ilr[, 1]	y_ilr[, 2]
Constant	$+1.01(0.91)$	$-2.35(0.89)^{**}$	$+1.34(7.90)^{***}$	$-1.48(6.60)^{***}$
Age_ilr1	$+0.05(0.78)$	$-0.53(0.76)$	$+0.17(6.76)^{***}$	$+0.44(5.64)^{***}$
Age_ilr2	$-0.35(0.45)$	$-0.75(0.44)^{*}$	$-0.44(3.96)^{***}$	$-0.94(3.31)^{***}$
unemp_rate	$-7.31(2.77)^{**}$	$+13.1(2.71)^{***}$	$-7.94(24.1)^{***}$	$+10.6(20.1)^{*}$
income_tax_rate	$-0.42(1.00)$	$+0.19(0.98)$	$-1.02(8.69)^{***}$	$-0.82(7.26)^{***}$
Note:				*p<0.1; **p<0.05; ***p<0.01

Bibliographie

Aitchison J. (1985). A General Class of Distributions on the Simplex, *Royal Statistical Society*, 47, pp. 136-146.

Ansolabehere S., Leblanc W. (2008). A spatial Model of the Relationship between Seats and votes, *Mathematical and Computer Modeling*, 48, pp. 1409-1420.

Beauguitte L., Colange C. (2013). Analyser les comportements électoraux à l'chelle du bureau de vote, ANR Cartelec : mémoire scientifique .

Chen, J., Zhang X., Li S. (2016). Multiple linear Regression with Compositional response and covariates, *Journal of Applied Statistics*, 44, pp. 2270-2285.

Egozcue, J.J., Pawlowsky-Glahn V., Mateu-Figueras G., Barcelo-Vidal C. (2003). Isometric Logratio Transformations for Compositional Data Analysis, *Mathematical Geology*, 35, pp. 279-300.

Johnson, N.L., Kotz S. (1972). Student Multivariate Distribution, In *Distributions in Statistics: continuous multivariate distributions*, Wiley, Michigan.

Honaker, J., Katz, J., King G. (2002). A Fast, Easy, and Efficient Estimator for Multiparty Electoral Data, *Political Analysis*, 10, pp. 84-100.

Katz, J., King G. (1999). A Statistical Model for Multiparty Electoral Data, *American Political Science Review*, 93, pp. 15-32.

Kelejian, H.H., Prucha, I.R. (1984). Independent or Uncorrelated Disturbances in Linear Regression, *Economics Letters*, 19, pp. 35-38.

Mert, M.C., Filzmoser, P., Endel, G. (2016). Compositional Data Analysis in Epidemiology, *Statistical Methods in Medical Research*, 27, pp. 1878-1891.

Morais, J., Thomas-Agnan, C. and Simioni, M. (2017) Using compositional and Dirichlet models for market share regression, Journal of Applied Statistics, al-01558527.

Nguyen, T.H.A., Laurent, T., Thomas-Agnan, C., Ruiz-Gazen, A. (2018). Analyzing the Impacts of Socio-Economic Factors on French Departmental Elections with CODA Methods, *TSE Working Paper*.

Nguyen, T.H.A., Ruiz-Gazen, A., Thomas-Agnan, C., Laurent, T. (2019). Multivariate Student versus Multivariate Gaussian Regression Models with Application to Finance, *Journal of Risk and Financial Management*, 12.