Copula Analysis of the 2018 Brazilian Presidential Election

J.A.A. Schultz[a)] and V.A. González-López[b)]

University of Campinas - Department of Statistics. Rua Sérgio Buarque de Holanda 651. Campinas, S.P. Brazil. CEP: 13083-859.

[a)]zeadolfo@gmail.com
[b)]veronica@ime.unicamp.br

Abstract. In this article, we stablish the dependence between the percentage of voters in the president of Brazil elected in the second round of 2018 and the percentage of self-declared evangelical voters. Considering the percentages state by state, we show that the dependence between both quantities can be well represented by a Gaussian copula, selected between six possible copulas. We identify the correlation coefficient ρ of the Gaussian copula through a Bayesian approach which allows us to determine a posterior distribution of ρ, with a mean value 0.727 and a 95% high density credibility interval $[0.531, 0.846]$.

Introduction

This work aims to investigate the relation between a specific social group of the population of Brazil and the abrupt change in the political line that governs Brazil since 2018. This research is especially relevant at this time when Brazil faces elections again, in October 2022, having as main actors in the electoral contest, the current president and to what is considered his opposite, the former president Lula da Silva. Under the democratic system, when drastic changes occur in the politics of a country, it could be a consequence of social changes happens in its population. For example, certain previously smaller groups may have become large enough to allow them to disrupt it previous profile. It appears to have been the situation faced by Brazil in 2018, when, after more than 15 years of center-left governments, it was elected a far-right candidate. In recent decades, one of the groups that have grown the most in the population of Brazil is evangelicals and it is for this reason that this is the group that we included in our inspection. With this goal, we investigate the relationship between the number of voters in the winner of the presidential elections in Brazil (2018) and the percentage of individuals who share the predominant religion in Brazil.

Data and Model Selection

We approach this problem using the concept of copula, since we want to model the relationship between X which is *the proportion of voters in the winner of the presidential elections in Brazil (2018)* and Y which is *the percentage of individuals who share the predominant religion in Brazil.* If H is the cumulative distribution function of (X, Y), there is a function C, such that $H(x, y) = C(F_X(x), F_Y(y))$, with $F_X(x) = \lim_{y\to\infty} H(x, y)$ and $F_Y(y) = \lim_{x\to\infty} H(x, y)$. And the function C is the 2-copula of (X, Y). $C(u, v) = \text{Prob}(F_X(X) \leq u, F_Y(Y) \leq v)$, for $u, v \in [0, 1]$, then, C is the distribution of the variables $U := F_X(X)$ and $V := F_Y(Y)$. To give flexibility to our analysis we investigate 6 types of dependences, (a) Gaussian copula, (b) t-student copula, (c) Frank copula, (d) Gumbel copula, (e) Clayton copula and (f) Joe copula (see [1]). The idea is to include a broad spectrum of dependence types since the t-student copula can fit better tail values and the archemedian copulas can detect better other kind of dependences. Then, U and V are represented by its pseudo-observations which the values u and v are related to the random variables U and V, build in our case by the empirical (and) marginal distributions, respectively.

The data is related to the 27 states of Brazil: Acre (AC), Alagoas (AL), Amapá (AP), Amazonas (AM), Bahia (BA), Ceará (CE), Distrito Federal (DF), Espírito Santo (ES), Goiás (GO), Maranhão (MA), Minas Gerais (MG), Mato Grosso (MT), Mato Grosso do Sul (MS), Pará (PA), Paraíba (PB), Pernambuco (PE), Paraná (PR), Piauí (PI), Rio de Janeiro (RJ), Rio Grande do Norte (RN), Rio Grande do Sul (RS), Rondônia (RO), Roraima (RR), Santa Catarina (SC), São Paulo (SP), Sergipe (SE), Tocantins (TO). The data set related to Y can be retrieved from `Instituto Brasileiro de Geografia e Estatística, Censo 2010`, https://censo2010.ibge.gov.br/apps/mapa [1], and the data set related to X is coming from `Tribunal Superior Eleitoral (TSE), October 7th 2018`, https://sig.tse.jus.br/ords/dwapr/seai/r/sig-eleicao-resultados/home?session=6044292194297 [2]

In order to proceed with the estimation of the model, the original observations $\{(x_i, y_i)\}_{i=1}^n$, where x_i is the proportion of evangelicals in the state i and y_i is the proportion of votes in Bolsonaro in 2018 in the second round in the state i, are replaced by their re-scaled marginal ranks to $[0,1]$ (pseudo-observations), $u_i := \frac{|\{j:1\leq j\leq n, x_j\leq x_i\}|}{n}$ and $v_i := \frac{|\{j:1\leq j\leq n, y_j\leq y_i\}|}{n}$, $i = 1, \cdots, n$, where $|A|$ denotes the cardinal of the set A, and $n = 27$. We see in Figure 1 the scatterplot between X and Y, represented by the observations $\{(x_i, y_i)\}_{i=1}^n$, and in Figure 2 the scatterplot between U and V, represented by the pseudo-observations $\{(u_i, v_i)\}_{i=1}^n$. Such figures (1, 2) support the idea that there is some dependence between X and Y, which becomes the focus of our investigation in the following lines. The Spearman's correlation coefficient given by the data is 0.7232, confirming the tendency exposed by the data in figure 2.

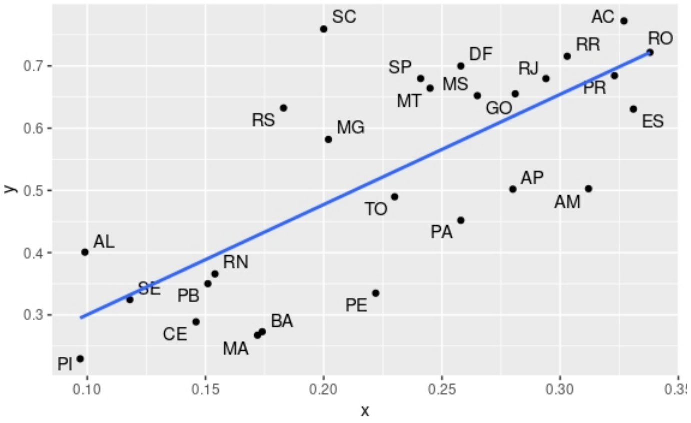

FIGURE 1. Scatterplot between the proportion of evangelicals (x) and proportion of votes for Bolsonaro (y), state by state.

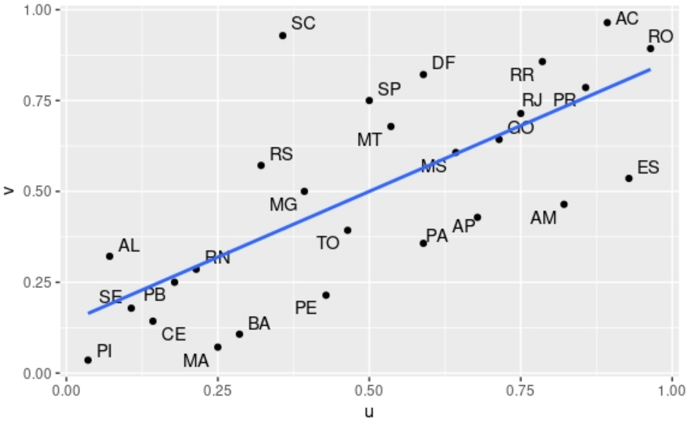

FIGURE 2. Scatterplot between pseudo-observations: u versus v, state by state.

Through the pseudo-observations $\{(u_i, v_i)\}_{i=1}^n$, for copula families (a)-(f) we construct the likelihood $\prod_{i=1}^n c(u_i, v_i)$,

[1] Last view, 17 of June of 2022

[2] Last view, 17 of June of 2022

where c is the density of the copula, in each case. In all the cases we use the *Copula R-package*, and the function *fitCopula()*, with arguments *copula* (1) and *method* (2), with (1) 'frankCopula(dim=2)', 'normalCopula(dim=2)', 'gumbelCopula(dim=2)', 'claytonCopula(dim=2)', 'tCopula(dim=2, t = 1)', 'tCopula(dim=2, t = 10)', 'joeCopula(dim=2)', respectively and (2) method= 'mpl'. The package also provides a goodness-of-fit test (see [2]) and it uses the maximum pseudo likelihood estimator to estimate the parameter for each copula. We report the results in Table 1.

TABLE 1. Goodness-of-fit for copula. In bold the indicated model.

Copula	**Gaussian**	Frank	t(10)	t(1)	Clayton	Gumbel	Joe
P-value	**0.934**	0.922	0.844	0.829	0.749	0.512	0.041

Given the results of Table 1, the Gaussian copula is the most suitable to represent the dependence between X and Y. Given the reduced number of data, $n = 27$, we implement a Bayesian approach to determine the correlation of the copula. This approach provides a flexible inspection, as we show in the next section.

Bayesian Inference on the Correlation Coefficient

Once fixed the Gaussian copula, we provide a Bayesian inspection of the correlation coefficient ρ, which is the copula's parameter in the case. For the study, we choose three prior distributions on ρ building three settings for ρ, (i) an impartial setting, with prior $\rho \sim$ Beta(1,1), (ii) partial setting, with prior $\rho \sim$ Beta(1,15), favoring negative dependence, (iii) a partial setting, with prior $\rho \sim$ Beta(15,1), favoring positive dependence. Note that since $\rho \in [-1, 1]$ the Beta distributions stated in i, ii and iii are Beta displayed in $[-1, 1]$ instead of $[0, 1]$.

Now we introduce briefly the Full Bayesian Significance Test (FBST), used as an alternative option to decide if we have or not significant evidence against the relation between X and Y. This is a test specially designed for a Bayesian context, see [3], [4]. Let $\pi\Big(\rho|\{(u_i, v_i)\}_{i=1}^n\Big)$ be the posterior distribution of ρ given $\{(u_i, v_i)\}_{i=1}^n$ and let be an evidence measure in $H_0 : \rho \in \Theta_0$ using the tangent set called T, with the more "probable" points for the set Θ_0,

$$T = \Big\{\rho \in \Theta : \pi\Big(\rho|\{(u_i, v_i)\}_{i=1}^n\Big) > t\Big\} \text{ where } t = \sup_{\rho\in\Theta_0} \pi\Big(\rho|\{(u_i, v_i)\}_{i=1}^n\Big).$$

Then, the evidence e-value in favor of the set Θ_0 (see [3]) is given by $e\text{-value} = 1-\mathbb{P}\Big(\rho \in T|\{(u_i, v_i)\}_{i=1}^n\Big)$. Small e-values point the rejection of H_0.

We estimate the posterior distributions (by Hamiltonian Monte-Carlo), under the settings (i), (ii) and (iii). We also report the e-value for $\Theta_0 = \{0\}$ (null coefficient), to decide if the hypothesis of independence between X and Y is feasible. The results are reported in Table 2.

TABLE 2. Summary measures for posterior distributions and independence test e-value

Case	Prior	Mín	1st Qu.	Median	3rd Qu.	Max	Mean	Std Dev	e-value
(i)	Beta(1,1)	0.173	0.684	0.742	0.785	0.896	0.727	0.083	< 0.001
(ii)	Beta(1,15)	-0.471	-0.048	0.063	0.187	0.712	0.072	0.172	0.353
(iii)	Beta(15,1)	0.381	0.735	0.775	0.811	0.892	0.769	0.060	< 0.001

Starting from the fact that the Spearman coefficient found in the data is 0.7232, cases (i) and (iii) are the most reasonable. Case (ii) (favorable to negative ρ values) is included to verify how it rearranges the ρ values when there is a conflict between the prior information and the likelihood function. (i) is adequate when there is no preference for any value of ρ in the support [-1,1], and case (iii) is adequate when the prior information tends to positive values of ρ. In such scenarios, the assumption of independence ($\rho = 0$) must be rejected since the e-values take negligible values (see the last column of table 2). We note that the impartial setting (i) indicates a Bayesian estimator, by quadratic loss function, equal to 0.727, that is, slightly higher than the observed correlation, while the most favorable scenario for

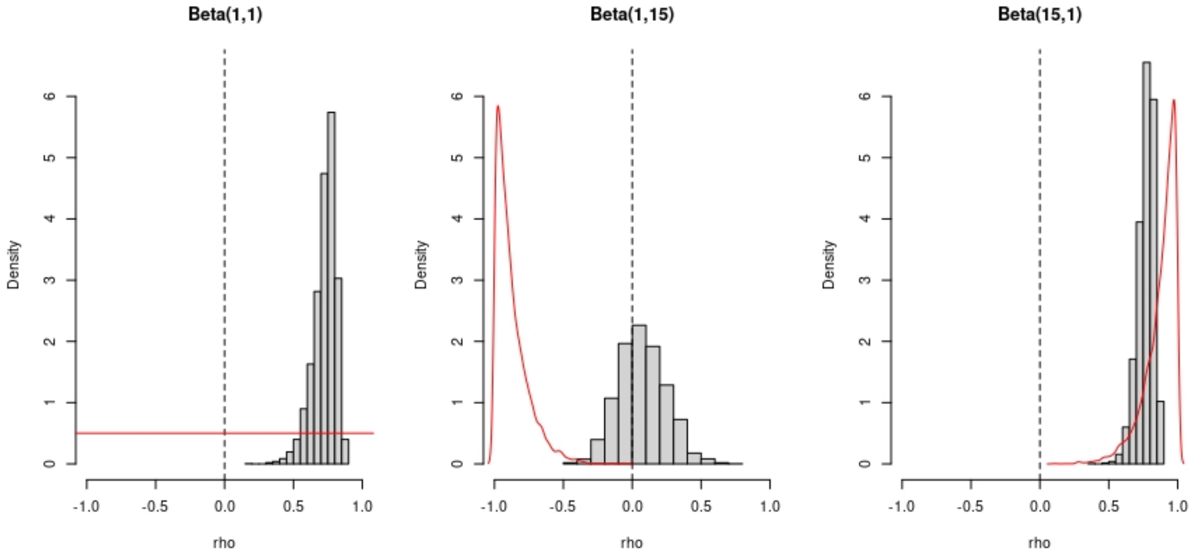

FIGURE 3. Histograms of posterior distributions with prior distributions Beta on ρ (in red). From left to right: setting i. (using a prior Beta(1,1)), setting ii. (using a prior Beta(1,15)) and setting iii. (using a prior Beta(15,1)).

positive correlation, (iii) gives a Bayesian estimator equal to 0.769, showing the effect of the prior distribution. We see in figure 3, when comparing (iii) with (i), that the posterior distribution of ρ suffers a (shy) reduction in its precision, 3rdQu-1stQu goes from 0.101 (in (i)) to 0.076 (in (iii)). On the other hand, for setting ii, the hypothesis $H_0 : \rho = 0$ is not rejected with e-value = 0.353. Moreover, the combination of the prior distribution (Beta(1,15)) and the likelihood function brings the posterior distribution around $\rho = 0$ (see figure 3-middle) showing a clear conflict between the data and the prior distribution.

Conclusions

In this work we identify the dependence between X which is *the proportion of voters in the winner of the presidential elections in Brazil (2018)* and Y which is *the percentage of individuals who share the predominant religion in Brazil*. We show that the dependence can be well represented by a Gaussian copula. To select the copula we use the test introduced in [2] (see table 1). In a second stage of this work, since the database is relatively moderate ($n = 27$), we conduct a Bayesian estimation of the Gaussian copula's correlation coefficient, see table 2 and figure 3. In this implementation we consider 3 settings that allow us to compare the effect of 3 prior distributions on the ρ correlation coefficient. The non-informative scenario (setting (i)) is taken as a reference and we confirm the non-nullity of such coefficient (ρ) by means of a genuinely Bayesian independence test, with e-value<0.001 (see [3], [4]). Also, this scenario gives a Bayesian estimator of $\rho = 0.727$ with a 95% high density credibility interval $[0.531, 0.846]$, then, we obtain a range of values for ρ confirming a non-irrelevant relation between X and Y.

REFERENCES

[1] R. B., Nelsen, *An introduction to copulas* (Springer, New York, 2007).

[2] C., Genest, B., Rémillard and D., Beaudoin, Goodness-of-fit tests for copulas: a review and a power study, *Insurance: Mathematics and Economics* **44**(2), 199-213 (2009). https://doi.org/10.1016/j.insmatheco.2007.10.005

[3] C.A.B., Pereira and J.M., Stern, Evidence and Credibility: Full Bayesian Significance Test for Precise Hypotheses. *Entropy* **1**(4), 99-110 (1999). https://doi.org/10.3390/e1040099

[4] M.R., Madruga, L.G., Esteves and S., Wechler, On the bayesianity of Pereira-Stern Tests. *Test* **10**(2), 291-299 (2001). https://doi.org/10.1007/BF02595698