Hierarchical Dirichlet regression model to benthic cover Abrolhos Bank.

Jéssica Areias, Larissa Maracajá, Mariana Soares Sá, Pamela M. Ch. Solano

Federal University of Rio de Janeiro - UFRJ

pamela@dme.ufrj.br

15TH Brazilian Meeting of Bayesian Statistics - Maresias Beach, São Sebastião - SP - Brazil, March 09-11, 2020

Intoduction

The Abrolhos area constitutes the largest and richest reef complex of the South Atlantic Ocean (Leitão et al., 2019). Learning about the dynamics of the Abrolhos bank involves ecological theory and environmental knowledge.

The components of benthic community are the proportions of a whole. This kind of data is denominated compositional data. The main feature is that their sum is constrained to one. This hidden structure of the Abrolhos bank is the interest of this contribution. A lot of examples of compositional data can be seen in Aitchison (2008).

The article developed by Brewer et al. (2005) includes a hierarchical structure in the model for compositional data via some distributions mixture. Furthermore, they included proportion values incorporating the zero and one in the analysis. Under frequentist inference Tsagris e Stewart (2018) fitted a Dirichlet regression allowing for zero values to be present in the analysis without any data modification.

Dirichlet Distribution

Let $\boldsymbol{y}_i = (y_{i1}, \ldots, y_{iC})'$ be a vector of proportions y_{ic}, where y_{ic} is defined as the proportion of coverage at observation $i = 1, \ldots, n$ of component $c = 1, \ldots, C$. The number of observations is n and C is the total number of components.

For each observation i, the random vector $\boldsymbol{Y}_i$ is distributed in $C \geq 2$ dimentions subject to $\sum_{c=1}^{C} y_{ic} = 1$. For that, the Dirichlet distribution is ideal.

The Dirichlet's probability density function is given by f($\boldsymbol{y}$):

$$f(\boldsymbol{y}) = \frac{\Gamma(\alpha_1 + ... + \alpha_c)}{\prod_{c=1}^{C} \Gamma(\alpha_c)} \prod_{c=1}^{C} y_c^{\alpha_c - 1}, \quad a_c, y_c > 0, \sum_{c=1}^{C} y_c = 1. \qquad (1)$$

The expected value for each dimension c of $\boldsymbol{Y}$ is $E[Y_c] = \frac{\alpha_c}{\alpha_0}$, the variance is $Var(Y_c) = \frac{\alpha_c(\alpha_0 - \alpha_c)}{\alpha_0^2(\alpha_0+1)}$ and the covariances $Cov(Y_c, Y_{c'}) = \frac{-\alpha_c \alpha_{c'}}{\alpha_0^2(\alpha_0+1)}$, where $\alpha_0 = \sum_{c=1}^{C} \alpha_c$.

Dirichlet regression under Bayesian approach

The basic Dirichlet regression model was originally proposed by Maier. We consider a Bayesian version of the basic regression to modeling the compositional data based on Holger e Sennhenn-Reulen (2018).

Following the notation in Maier, if we define $\mu_c = E(Y_c) = \frac{\alpha_c}{\phi}$ to account for the expected values of the variables Y_c where $\phi = \alpha_0 = \sum_{c=1}^{C} \alpha_i$, we have a parameterization $Y \sim D(\boldsymbol{\mu}, \phi)$. To convert μ and ϕ back to the Dirichlet distribution's original $\boldsymbol{\alpha}$ parameters we define $\alpha_c = \mu_c \phi$. Aditionally, we can introduce the $E(Y_c) = \mu_c$, $V(Y_c) = \frac{\mu_c(1-\mu_c)}{\phi+1}$, and $Cov(Y_c, Y_{c'}) = \frac{-\mu_c \mu_{c'}}{\phi+1}$, with $\mu_c \in (0, 1)$ and $\phi > 0$. Regressors can be introduced by making

$$\boldsymbol{Y}_i \mid \boldsymbol{x}_i \sim D(\boldsymbol{\mu}_i, \phi_i), \qquad (2)$$

where $\boldsymbol{x}_i$ is a P-dimensional vector of regressors. The expectation vector becomes $\boldsymbol{\mu}_i = (\mu_{i1}, \ldots, \mu_{iC})'$ where $\mu_{ic} = E(Y_{ic} \mid x_i)$ is the c^{th} element in the expectation vector and precision ϕ_i.

Equation 2 introduces regressors into the model. The component-wise P-dimensional coefficients vectors $\boldsymbol{\beta}_c$ for components $c = 1, \ldots, C$ and the covariates define the linear predictors $\eta_{ic} = \boldsymbol{x}_i' \boldsymbol{\beta}_c$. This regression is on the expectation vector $\boldsymbol{\mu}_i$ thus each component can be written as

$$\mu_{ic} = E(Y_{ic} \mid \boldsymbol{x}_i) = \frac{\exp(\boldsymbol{x}_i' \boldsymbol{\beta}_c)}{\sum_{d=1}^{C} \exp(\boldsymbol{x}_i' \boldsymbol{\beta}_d)} \qquad (3)$$

Note that identificability needs to be ensured as a consequence of the degrees of freedom being reduced to $C - 1$. For this purpose, a Δ reference component should be chosen and $\boldsymbol{\beta}_\Delta$ is set to zero. Usually Δ is selected as 1 or C, and in our aplications the C$^{\text{th}}$ component was chosen. Let $\Theta = (\boldsymbol{\beta}, \phi)$ be the vector of parameters. Using $\alpha_c = \mu_c \phi$, we obtain the density

$$f(\boldsymbol{y}_i \mid \boldsymbol{\Theta}) = f(\boldsymbol{y}_i \mid \boldsymbol{\mu}_i, \phi) = \frac{\Gamma(\sum_{c=1}^{C} \mu_{ic}\phi)}{\prod_{c=1}^{C} \Gamma(\mu_{ic}\phi)} \prod_{c=1}^{C} y_c^{\mu_{ic}\phi - 1}.$$

The Δ reference component is given by $\mu_{i\Delta} = \frac{1}{\sum_{d=1}^{C-1} \mu_{id}}$.

Inference: Posterior sampling

The likelihood $L(\boldsymbol{\Theta} \mid \boldsymbol{y}_i; \quad i = 1, \ldots, n) = \prod_{i=1}^{n} f(\boldsymbol{y}_i \mid \boldsymbol{\Theta})$ is derived.

$$L(\boldsymbol{\Theta} \mid \boldsymbol{y}) = \sum_{i=1}^{n} (\log \Gamma(\sum_{c=1}^{C} \mu_{ic}\phi) - \sum_{c=1}^{C} \log \Gamma(\mu_{ic}\phi) + \sum_{c=1}^{C} (\mu_{ic}\phi - 1) \log(p_c)).$$

where the μ_{ic} are functions of $\boldsymbol{\beta}_c$ which are unknown components to be estimated. The vector $\boldsymbol{y} = (\boldsymbol{y}_1, \ldots, \boldsymbol{y}_n)$ denotes all the information available provided by the data.

To complete the model specification we assigned an appropriate proper prior distribution to the parametric vector $\boldsymbol{\Theta}$. Prior independence between the parameters is assumed to the model parameters. The choice of the independent prior distributions were driven for the need of making inference with minimum subject prior information. The parameter $\theta = \log(\phi)$ and $\beta_{cp}, \quad c = 1, \ldots, C, \quad p = 1, \ldots, P$ are normally distributed with zero mean and precision $1/K$ for all effects of the model.

Assuming independence in the prior distributions, the posterior distribution is given by

$$\pi(\boldsymbol{\Theta} \mid \boldsymbol{y}) \propto L(\boldsymbol{\Theta} \mid \boldsymbol{y}) \pi(\phi) \prod_{c,p}^{C,P} \pi(\beta_{cp}) \qquad (4)$$

Since the joint posterior distribution in (4) does not have a known closed form, we propose the use of MCMC methods to obtain samples from it. Sampling from the distribution whose density is in equation (4) is done by Markov chain Monte Carlo (MCMC) using the No-U-Turn-Sampler implemented in Stan software.

Aplication: Abrolhos bank

The Abrolhos Bank is a $46.000 km^2$ extension of the continental shelf off Bahia and Espírito Santo states, eastern Brazil. The region is characterized by a complex mosaic of benthic megahabitats that encompass the largest $(8844 km^2)$ and richest reefs in the tropical southwestern Atlantic Ocean (SWA), and the world's largest rhodolith beds (Leitão et al., 2019). Corals under each point were identified and visually assigned in C categories.

The components of benthic community are the percentages of a whole. The data is representated in Figure1.

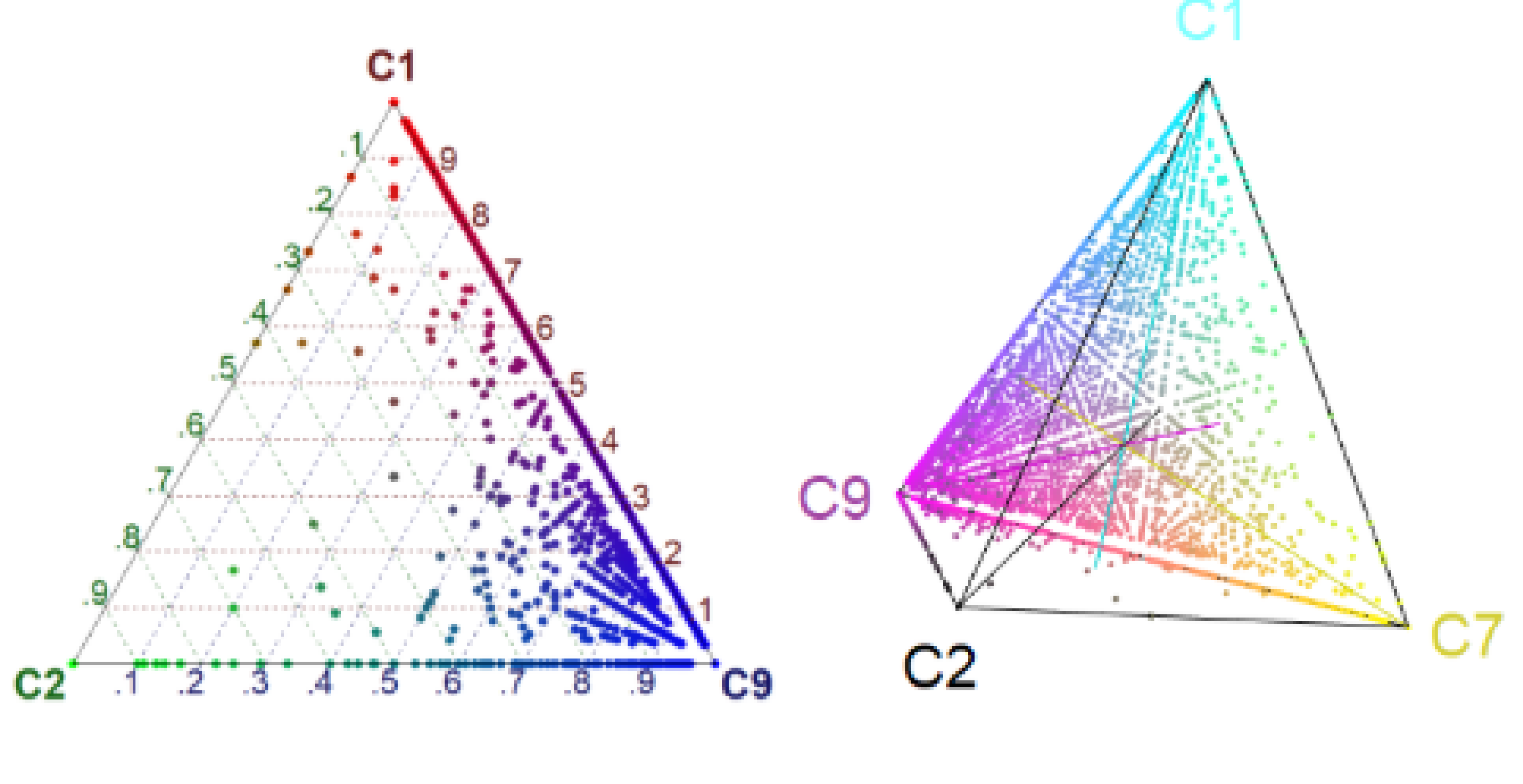

Figure 1: *Example for four components (C1-C2-C7-C9) of benthic community*

The main goal is to analyze the effects β_{cp} of each dummy variable which represent each Site $p = 1, \ldots, 6$ in each component $c = 1, \ldots, 9$ (Coral, Fire Coral, Sponge, Zoan, CCA, Cyano, Macro, Turf and others). The Δ component TURF was selected as reference. There were a total of $n = 19909$ observations. The $\boldsymbol{x}_i$ is a $P = 6$-dimensional vector which represents the location where the sample was collected. The model described in Model Section was used and the independents normal prior distributions for the parameteres $\boldsymbol{\beta}, \phi = \exp(\theta)$ were as defined in Inference Section.

Four Markov chains of length of 10,000 each, starting from different starting points with warm-up of 5000 iterations were generated. Convergence was monitored via MCMC chain trajectories, autocorrelation, cross-correlation and density plots by use of the package coda Plummer et al. (2006), available in R software.

After fitting under the Bayesian approach, we get the coefficient estimates using the frequentist approach, for this the data is analyzed using the DirichReg function based on Maier. Figure 2 compares this results.

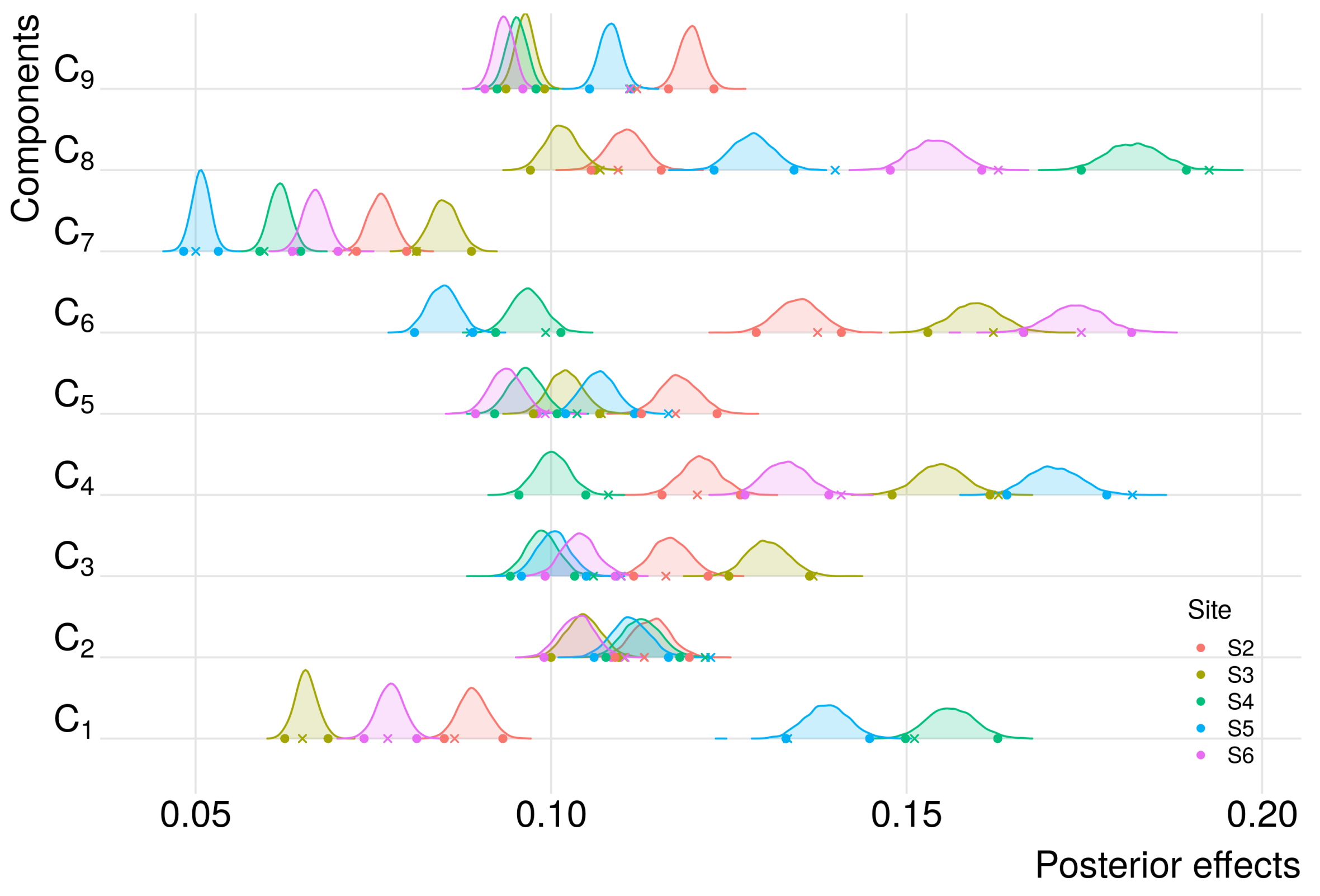

Figure 2: *Density of the marginal posterior of the $\boldsymbol{\mu}$ effects for each of the nine components. The circles describe the credible intervals and the "×" shows the estimate under the frequentist approach.*

Figure 2 indicates that in all sites, except for S3, the effects over C7 are smaller. On the other hand, the effects on sites S2, S3, S5 and S6 are bigger over C6 while the effect of all sites over C2 is homogeneous.

In almost all of the sites the frequentist estimation is inside the credible interval. That is a good sign because it shows that the Bayesian results match the frequentist ones, whose theory is very well consolidated. The only component that is outside the credible interval in all sites is the C9 one.

Future work

The model applied to the Abrolhos bank was defined based on the expected values of the variables Y_c. The constraint $\boldsymbol{\beta}_\Delta = 0$, provides diferent interpretations depending which Δ reference component is selected. Covariates such as environmental variables can be incorporated to obtain a more realistic model.

References

Aitchison, J. (2008) The statistical analysis of compositional data. **44**, 139–177.

Brewer, M., Filipe, J. e Elston, D. (2005) A hierarchical model for compositional data analysis. URL`https://doi.org/10.1198/108571105X2820`.

Holger e Sennhenn-Reulen (2018) Bayesian regression for a dirichlet distributed response using stan. URL`https://arxiv.org/abs/1808.06399`.

Leitão, R., Ribeiro, F., Solano, P., Teixeira, C., Magdalena, U., Salomon, P., Villella, L., Bastos, A., Falsarella, L., Cardoso, G., Pereira-Filho, G., Salgado, L., L., N. e Moura, R. (2019) *Interspecific variation in resilience after mass Coral beaching in southwestern atlantic turbid zone reefs*. Dissertação de Mestrado.

Maier, M. J. () Dirichletreg: Dirichlet regression for compositional data in r. URL`http://statmath.wu.ac.at/`.

Plummer, M., Best, N., Cowles, K. e Vines, K. (2006) Coda: Convergence diagnosis and output analysis for mcmc. *R News*, **6**, 7–11.

Tsagris, M. e Stewart, C. (2018) A dirichlet regression model for compositional data with zeros. **39**, 398—-412.