Estimation and Model Selection in Dirichlet Regression

André P. Camargo*, Julio M. Stern* and Marcelo S. Lauretto†,**

*University of Sao Paulo, Institute of Mathematics and Statistics
†University of Sao Paulo, School of Arts, Sciences and Humanities
**marcelolauretto@usp.br

Abstract. We study Compositional Models based on Dirichlet Regression where, given a (vector) covariate $\mathbf{x}$, one considers the response variable $\mathbf{y} = (y_1, \dots, y_D)$ to be a positive vector with a conditional Dirichlet distribution, $\mathbf{y}|\mathbf{x} \sim \mathscr{D}(\alpha_1(\mathbf{x}) \dots \alpha_D(\mathbf{x}))$. We introduce a new method for estimating the parameters of the Dirichlet Covariate Model when $\alpha_j(\mathbf{x})$ is a linear model on $\mathbf{x}$, and also propose a Bayesian model selection approach. We present some numerical results which suggest that our proposals are more stable and robust than traditional approaches.

Keywords: Dirichlet regression, parameter estimation, model selection
PACS: 01.30.Cc, 02.50.Ng, 02.50.-r

INTRODUCTION

Compositional data consist of vectors whose components are the proportions of some whole. The peculiarity of these models is that the sum of the sample vector adds to 1. Hence, the corresponding sample space is the simplex, that is quite different from the real Euclidean space associated with unconstrained data. Therefore, attempts to apply statistical methods for unconstrained data often lead to inappropriate inference. Some statistical models for compositional data have been developed since the 70s, particularly for regression analysis. Aitchison [1, 2] exploited the logistic normal distribution properties and developed a complete theory of compositional data analysis, based on some classes of logistic transformations from the real space to the simplex.

Here we focus on a less exploited approach, the Dirichlet Covariate Model, suggested by Campbel and Mosimann [7]. In this model, one considers $\mathbf{y} = (y_1, \dots, y_D)$ to be a $1 \times D$ positive vector having Dirichlet distribution $\mathscr{D}(\alpha_1, \dots, \alpha_D)$. A Dirichlet regression model is readily obtained by allowing its parameters to change with covariates. For a given covariate row vector $\mathbf{x} = (x_1, \dots, x_C)$, each parameter α_j may be written as a positive-valued vector function $\alpha_j(\mathbf{x}), j = 1 \dots D$ and, therefore, the response variable is assumed to follow a conditional Dirichlet distribution, $\mathbf{y}|\mathbf{x} \sim \mathscr{D}(\alpha_1(\mathbf{x}) \dots \alpha_D(\mathbf{x}))$.

We focus on the case where each $\alpha_j(\mathbf{x})$ is a linear function of $\mathbf{x}$. This formulation is intuitive and resembles, in some aspects, standard linear models interpretation. However, the constraint $\alpha_j(\mathbf{x}) > 0$ may impose some difficulties for estimation and inference methods. We propose a new method for parameter estimation, which outperforms the current approach [9], and propose a model selection approach based on the Full Bayesian Significance Test [12, 13].

Bayesian Inference and Maximum Entropy Methods in Science and Engineering
AIP Conf. Proc. 1443, 206-213 (2012); doi: 10.1063/1.3703637

DIRICHLET REGRESSION

Let $\mathscr{S}^D$ denote the $(D-1)-$dimensional simplex:

$$\mathscr{S}^D = \{\mathbf{z} = (z_1, z_2 \dots z_D) : \mathbf{z} > 0, \sum_{j=1}^{D} z_j = 1\}.$$

The vectors $\mathbf{0}$ and $\mathbf{1}$ always have the appropriate dimension required by the context. Let $X = (\mathbf{x}_{1\bullet}; \mathbf{x}_{2\bullet}; \dots; \mathbf{x}_{n\bullet})$, $Y = (\mathbf{y}_{1\bullet}; \mathbf{y}_{2\bullet}; \dots; \mathbf{y}_{n\bullet})$ be a sample of observations where $\mathbf{y}_{i\bullet} \in \mathscr{S}^D$ and $\mathbf{x}_{i\bullet} \in R^C$, $i = 1, 2, \dots, n$. The goal is to build a regression predictor for $\mathbf{y}_{i\bullet}$ as a function of $\mathbf{x}_{i\bullet}$.

We assume that $\mathbf{y}_{i\bullet}$ follows a Dirichlet distribution with parameter $\alpha(\mathbf{x}_{i\bullet})$, where $\alpha(\mathbf{x}_{i\bullet}) = (\alpha_1(\mathbf{x}_{i\bullet}), \dots, \alpha_D(\mathbf{x}_{i\bullet}))$, and each $\alpha_j(\mathbf{x}_{i\bullet})$ is a linear combination of $\mathbf{x}_{i\bullet}$:

$$\alpha_j(\mathbf{x}_{i\bullet}) = x_{i,1}\beta_{1,j} + x_{i,2}\beta_{2,j} + \dots + x_{i,C}\beta_{C,j} = \mathbf{x}_{i\bullet}\beta_{\bullet j}.$$

The parameters to be estimated are $\beta = (\beta_{k,j}, k = 1 \dots C, j = 1 \dots D)$, subject to the constraint $\alpha(\mathbf{x}_{i\bullet}) > 0$. As usual, model selection can be done by testing $\beta_{k,j} = 0$ for some pairs $(k, j) \in \{1 \dots C\} \times \{1 \dots D\}$.

Parameter Estimation

Assuming that $\mathbf{y}_{1\bullet} \dots, \mathbf{y}_{n\bullet}$ are c.i.i.d. given β, the likelihood function is:

$$L(\beta \,|\, X, Y) = \prod_{i=1}^{n} \left[\Gamma(A_i(\mathbf{x}_{i\bullet})) \prod_{j=1}^{D} \frac{y_{ij}^{\alpha_j(\mathbf{x}_{i\bullet})-1}}{\Gamma(\alpha_j(\mathbf{x}_{i\bullet}))} \right],$$

where $A_i(\mathbf{x}_{i\bullet}) = \sum_{j=1}^{D} \alpha_j(\mathbf{x}_{i\bullet})$.

In order to make fair comparisons with frequentist mehods, we use in this article the uniform (improper) prior for β, and the last expression is also the posterior distribution. In a more general setting, a Bayesian user may choose to use a more adequate (proper) prior.

The gradients of log-likelihood are easily computed, and used for maximum likelihood estimation:

$$\frac{\partial \log L}{\partial \beta_{k,j}} = \sum_{i=1}^{n} \left[\left(\psi(A_i(\mathbf{x}_{i\bullet})) - \psi(\alpha_j(\mathbf{x}_{i\bullet})) + \log y_{i,j} \right) x_{i,k} \right]$$

where ψ denotes the digamma function, $\psi(u) = \frac{\partial \log \Gamma}{\partial u}(u)$.

Numerical methods are required for computing the Maximum Likelihood Estimates (MLE). Fitting a Dirichlet Distributions with constant parameters is straightforward, and numerical packages are available [5, 4]. The difficulty arises when we attempt to extend the estimation to Dirichlet Regression. Starting values and regularization policies must be carefully chosen for the optimization algorithm to converge.

Hijazi and Jernigan [9] proposed the following method for choosing starting values for the optimization step:

1. Draw r samples with replacement each of size m $(m < n)$ from X and Y.

2. For each sample, fit a Dirichlet model with constant parameters, and compute the mean of the corresponding covariates. This will result in matrices $A\ \ r \times D, W\ \ r \times C$ where A represents the ML estimates for the r samples and row $\mathbf{w}_i$ represents the means of covariates in sample i.
3. Fit by least squares D models of the form $A_{i,j} = \alpha_j(\mathbf{w}_i) = \sum_{k=1}^{C} \beta_{jk} w_{ik}$.
4. Use the fitted coefficients $\beta_{k,j}$ as starting values.

The main issue of Hijazi's method is that it does not guarantee the starting values $\beta_{k,j}$ to yield positive values for $\alpha_j(x_i)$. Taking this issue in account, our implementation for this method repeats the steps 1-4 until a feasible initial guess is obtained, limited to 30 uncussessfull iterations.

We propose a regularization approach anchored by the constant (without covariates) Dirichlet model. If the initial model does not include the constant (intercept) terms, we extend the initial model to include them as artificial variables. Finally, we solve a sequence of optimization problems that drive the artificial variables back to zero. The algebraic formulation of this procedure is as follows.

1. Add a constant column $\mathbf{1}$ as the first column of X, in case such an "intercept" column is not already present in the original model.
2. Define a boolean matrix M indicating the non-zero parameters of the original model, namely:
$$M_{k,j} = \begin{cases} 1 \text{ if } \beta_{k,j} \text{ is a model parameter;} \\ 0 \text{ if } \beta_{k,j} = 0. \end{cases}$$
3. Fit a Dirichlet model for Y with constant parameters (via MLE).
Notice that this corresponds to the solution of a basic model where the M^0 is
$$M^0_{k,j} = \begin{cases} 1 \text{ if } k = 1 \\ 0 \text{ if } k \neq 1 \end{cases}$$
Moreover, this solution is a feasible point for the (possible extended) model including the intercept.
4. Build the supermodel joining all variables present either in the anchor or in the original model, namely:
$$M^*_{k,j} = \max(M^0_{k,j}, M_{k,j}), k = 1 \ldots C, j = 1 \ldots D.$$
5. Solve the sequence of optimization problems
$$\max_{\beta} g(\beta \,|\, X, Y) = -K * \mathbf{b}\beta^2 + \log L(\beta \,|\, X, Y).$$
The boolean vector $\mathbf{b}$ indicates which of the $\beta_{1,j}$ are "artificial" variables (1) and which were present in the original model (0):
$$b_j = \begin{cases} 1 \text{ if } M_{j,1} = 0; \\ 0 \text{ otherwise.} \end{cases}$$
A sequence of increasing scalars, K_t, steadily increases the importance of the penalty term in the objective function. Each solution in the sequence is used as

the starting point for the next optimization problem. The increasing penalty term drives the artificial variables to zero, converging to the optimal solution (best fit) of the original model.

Taking a gradually increasing sequence of the penalty constants, K_t, constitutes what, in the optimization literature, is called a regularization policy [10]. Its use facilitates the convergence of the process to the desired optimum solution. Numerical experiments demonstrate that the proposed procedure is more stable and faster than Hijazi's approach. A case study is presented in the Results Section.

Prediction Using Dirichlet Regression

Having obtained the estimate $\hat{\beta}$, the expected composition proportions in $\mathbf{y}$ given the vector $\mathbf{x}$ of covariates values is the mean of the distribution $\mathscr{D}(\hat{\alpha}(\mathbf{x}))$:

$$\hat{\mathbf{y}} = \left(\frac{\hat{\alpha}_1(\mathbf{x})}{\hat{A}(\mathbf{x})}, \frac{\hat{\alpha}_2(\mathbf{x})}{\hat{A}(\mathbf{x})} \dots \frac{\hat{\alpha}_D(\mathbf{x})}{\hat{A}(\mathbf{x})} \right), \quad \text{where } \hat{A}(\mathbf{x}) = \sum_{j=1}^{D} \hat{\alpha}_j(\mathbf{x}).$$

FULL BAYESIAN SIGNIFICANCE TEST (FBST)

The Full Bayesian Significance Test (FBST) is presented in [12, 13] as a coherent Bayesian significance test. FBST is suitable for cases where the parameter space, Θ, is a subset of R^n, and the hypothesis is defined as a restricted subset defined by vector valued inequality and equality constraints: $H : \theta \in \Theta_H$, where $\Theta_H = \{\theta \in \Theta \,|\, g(\theta) \leq 0 \wedge h(\theta) = 0\}$. For simplicity, we often use H for Θ_H. We are interested in precise hypotheses, with $\dim(H) < \dim(\Theta)$. In this work, $f_x(\theta)$ denotes the posterior probability density function.

The computation of the evidence measure used on the FBST is performed in two steps:

- The *optimization step* consists of finding the maximum (supremum) of the posterior under the null hypothesis, $\theta^* = \arg\sup_H f_x(\theta), \ f^* = f_x(\theta^*)$.
- The *integration step* consists of integrating the posterior density over the Tangential Set, $\overline{T}$, where the posterior is higher than anywhere in the hypothesis, i.e.,

$$\overline{T} = \{\theta \in \Theta : f_x(\theta) > f^*\} \ , \quad \overline{\mathrm{Ev}}(H) = \Pr(\theta \in \overline{T} \,|\, x) = \int_{\overline{T}} f_x(\theta) d\theta$$

$\overline{\mathrm{Ev}}(H)$ is the evidence against H, and $\mathrm{Ev}(H) = 1 - \overline{\mathrm{Ev}}(H)$ is the evidence supporting (or in favor of) H. A more detailed FBST review may be found in [13].

In this work, θ corresponds to the model coefficients, i.e, $\theta = (\beta_{k,j}, j = 1 \dots D, k = 1 \dots C)$. For FBST implementation, we assume an improper uniform prior on $R^{D \times C}$, and therefore $f_x(\beta) \propto L(\beta \,|\, X, Y)$. For numerical integration, we adopt a Metropolis-Hastings algorithm with a multivariate normal proposal distribution [11, 15]. We use an initially diagonal kernel estimated at the maximum likelihood point, $\hat{\beta}$. The kernel is then periodically updated using standard adaptive methods [16]Ap.G.

RESULTS AND FINAL COMMENTS

This section presents some numerical experiments motivated by the benchmark application *Arctic lake sediments*, presented by Coakley and Rust [8] and adapted by Aitchison [2]. It consists of compositional data of sand, silt and clay for 39 sediment samples at different water depths. The immediate questions are [2]:

1. Is sediment composition dependent on water depth?
2. If so, how can we quantify this dependence?

Hence, the dataset comprises a response matrix Y of order 39×3 (constituents proportions in each sample), while the covariate matrix starts as a matrix X of order 39×1 (the sample depth).

The models of interest in our study are submodels of the complete second-order polynomial model on x,

$$\alpha_j(x) = \beta_{1,j} + \beta_{2,j} * x + \beta_{3,j} * x^2,\ j = 1 \ldots 3.$$

Figure 1 shows the Arctic Lake dataset with the corresponding complete first and second order models (continuous and dashed curves, respectively).

Parameter Estimation Procedures

In order to evaluate our proposed method for estimating polynomial coefficients, we draw a collection of $q = 1, 2, \ldots$ subsamples of the Arctic Lake dataset. We used subsamples with $50\%(n = 20)$ and $70\%(n = 27)$ of the points in the original dataset.

We try to fit each subsample with an incomplete polynomial model described by a random structural matrix $M^{(q)}$. The elements of each structural matrix, $M^{(q)}_{k,j}$, are set by a Bernoulli process where $Pr(M^{(q)}_{k,j} = 1) = p$. The fill-in probability, p, of setting to 1 an element of $M^{(q)}$ was set to $p = 0.33$, $p = 0.5$ or $p = 0.66$. Moreover, in order to avoid inconsistent models, a structural matrix is rejected if it does not satisfy the feasibility constraints $\sum_{k=1}^{C} M_{k,j}(q) \geq 1, j = 1 \ldots D$. These constraints make sure that there is at least one non-zero polinomial coefficient for each parameter function, $\alpha_j(x)$.

For each pair $(n, p) \in \{20, 27\} \times \{0.33, 0.5, 0.66\}$, we draw $m = 1000$ subsamples.

Two performance measures were considered: (1) The failure rate in the numerical optimization process used to fit the model; (2) The computational processing time of each method. Figure 2 presents the failure rate (left) and the processing time in Log$_2$ scale (right), according to the model fill-in probability. It is clear that both the failure rate and the processing time of Hijazi's method are much higher than ours. It is also clear that the performance of Hijazi's method deteriorates as the fill-in probability (structural density) increases, whereas our method does not seem to be affected. The subsample size ($n = 20$ or $n = 27$) does not seem to affect the performance. Hence, the experiments were pooled together.

Hypothesis Tests

In this section, we compare the performance of the FBST, as presented in Section 3, and the Likelihood Ratio (LR) test, a standard and easily computed classical approach. Performance comparison is based on the analysis of Type I and Type II errors (respectively,

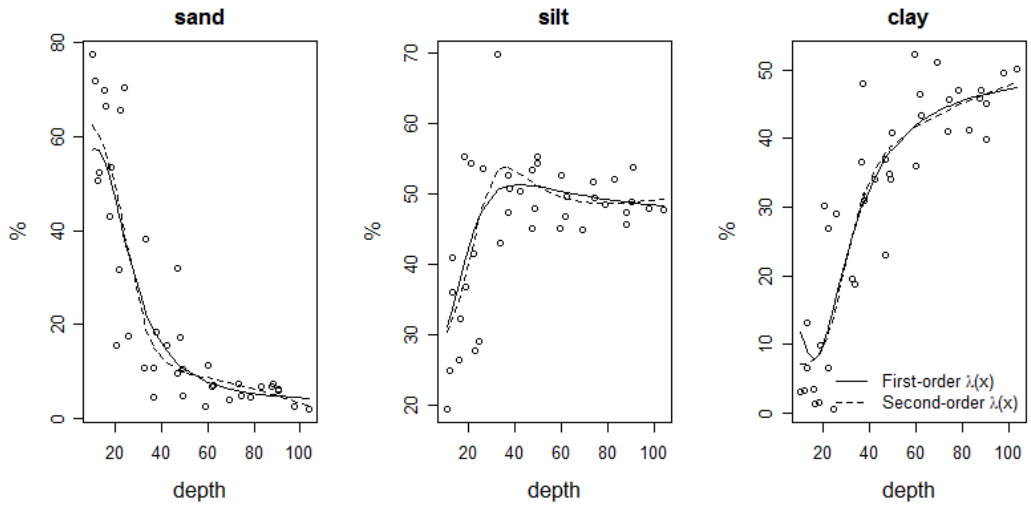

FIGURE 1. *Arctic Lake* dataset and corresponding fitted models: first order (continuous curves) and second order (dashed curves) polynomials

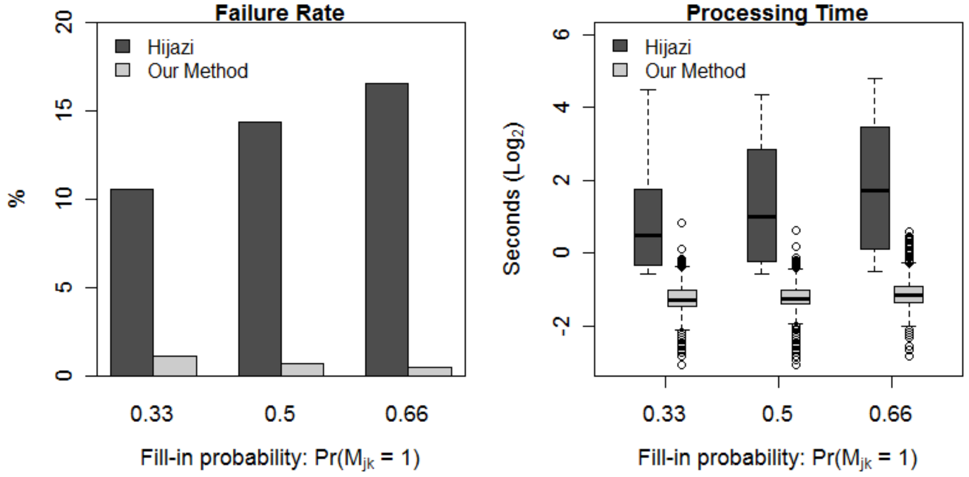

FIGURE 2. Failure rates (left) and log-processing times (right), by model fill-in probability.

the rejection rate of a true hypothesis and the acceptance rate of a false hypothesis). As it is standard in the literature, the basic idea is to compare the performance of Type II errors after setting acceptance/rejection thresholds τ_{FBST} and τ_{LR}, corresponding to an expected Type I error of $\alpha = 5\%$, that is, a $(1-\alpha) = 95\%$ confidence level. In this work, we consider two approaches for establishing τ_{FBST}:

(1) Asymptotic approximation: an asymptotic approximation to the threshold τ_{FBST} can be easily computed, as explained in [6, 13].

(2) Empirical power analysis: an empirical approximation to the thresholds τ_{FBST} and τ_{LR} can be obtained by computational simulation, as explained in [3].

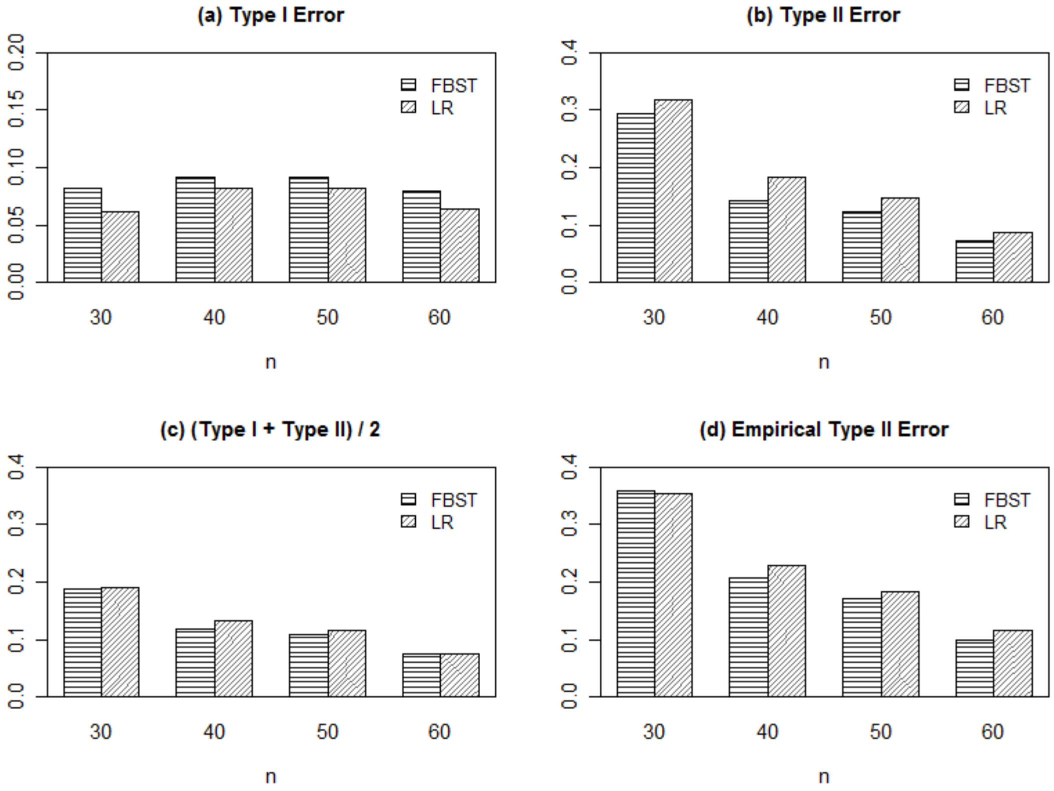

FIGURE 3. (a) Type I error, (b) Type II error and (c) average errors with asymptotic thresholds τ_{FBST} and τ_{LR}; (d) Type II errors with empirical thresholds τ_{FBST} and τ_{LR}.

As in the last section, we start with the complete second-order polynomial model and use, as a benchmark hypothesis for the numerical experiments, $H : \beta_{3,j} = 0, j = 1,2,3$, that is, the assumption that $\alpha_j(x)$ may be suitable modelled as a first-order polynomial. Let θ^* and $\hat{\theta}$ represent the constrained (first order polinomial) and unconstrained (second order polinomial) maximum likelihood parameters optimized to the Arctic Lake dataset.

For each sample size $n \in \{30,40,50,60\}$, we generate two collections of $m = 500$ independent samples of size n. The first collection, $\mathscr{C}_1$, consists of samples drawn with parameter θ^*. The second collection, $\mathscr{C}_2$, consists of samples drawn with parameter $\hat{\theta}$. Type I errors of FBST and LR are estimated as the proportion of samples in $\mathscr{C}_1$ wich evidence in favour of H is less than τ_{FBST} and τ_{LR}, respectively. Type II errors of FBST and LR are estimated as the proportion of samples in $\mathscr{C}_2$ wich evidence in favour of H is greater than τ_{FBST} and τ_{LR}, respectively.

Figure 3(a,b) presents the Type I and Type II errors for the FBST and the LR, estimated with the asymptotic thresholds τ_{FBST} and τ_{LR}. Figure 3(c) presents the average errors, i.e., (Type I error + Type II error)/2. Although FBST has a slightly higher Type I error than LR, it achieves lower Type II and average errors than LR test.

Figure 3(d) presents the estimated Type II error, based on the empirical thresholds τ_{FBST} and τ_{LR}. FBST empirical Type II error is lower than LR test, suggesting a better discriminant power for FBST.

Future works shall compare the FBST performance with other model selection approaches not mentioned in this study, like AIC, BIC, etc. The routines used in this paper were developed on R and are available upon request. In the future, we hope to make these routines available as CRAN R packages [14].

Acknowledgments

The authors are grateful for the support of the School of Arts and Sciences (EACH) and Institute of Mathematics and Statistics (IME) of the University of São Paulo, CAPES - Coordenação de Aperfeiçoamento de Pessoal de Nível Superior, CNPq - Conselho Nacional de Desenvolvimento Científico e Tecnológico, and FAPESP - Fundação de Amparo à Pesquisa do Estado de São Paulo.

REFERENCES

1. Aitchison, J. (1982). The statistical analysis of compositional data (with discussion). *Journal of the Royal Statistical Society*, Series B, 44, 139-177.
2. Aitchison, J.(1986) The Statistical Analysis of Compositional Data, Monographs on Statistics and Applied Probability (Chapman & Hall Ltd, London) Reprinted (2003) with additional material by The Blackburn Press, Caldwell, NJ.
3. Bernardo, B.B., Lauretto, M.S., Stern, J.M. (2011) The Full Bayesian Significance Test for Symmetry in Contingency Tables. To be published in this volume.
4. Birgin, E.G, Castillo, R., Martinez, J.M.(2004). Numerical comparison of Augmented Lagrangian algorithms for nonconvex problems. *Computational Optimization and Applications* 31(1), 31-55.
5. Boogaart, K.G.v.d., Tolosana-Delgado, R. (2008): "compositions": a unified R package to analyze compositional data. *Computers & Geosciences* 34(4), 320-338.
6. Borges, W., Stern, J.M. (2007). The Rules of Logic Composition for the Bayesian Epistemic e-Values. Logic J.of the IGPL, 15, 5-6, 401-420
7. Campbell, G. , and Mosimann, J. (1987). Multivariate methods for proportional shape. *ASA Proceedings of the Section on Statistical Graphics*, 10-17.
8. Coakley, J.P. and Rust, B.R. (1968). Sedimentation in an Arctic lake. *Sedimentary Petrology* 38, 1290-1300.
9. Hijazi, R.H., and Jernigan, R.W. (2009). Modelling Compositional Data Using Dirichlet Regression Models. *Journal of Applied Probability & Statistics* 4, 77-91.
10. Lauretto, M.S., Pereira, C.A.B., Stern, J.M. (2003). Full Bayesian significance test applied to multivariate normal structure models. *Brazilian Journal of Probability and Statistics* 17, 147-168.
11. Martin, A.D., Quinn, K.M., Park, J.H. (2011). MCMCpack: Markov Chain Monte Carlo in R. *Journal of Statistical Software* 42(9): 1-21. `http://www.jstatsoft.org/v42/i09/`
12. Pereira, C.A.B., Stern, J.M. (1999). Evidence and Credibility: Full Bayesian Significance Test for Precise Hypotheses. *Entropy,* 1, 99-110.
13. Pereira, C.A.B., Stern, J.M., Wechsler, S. (2008). Can a significance test be genuinely Bayesian? *Bayesian Analysis* 3(1), 79-100.
14. R Development Core Team (2011). *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing, Vienna, Austria. http://www.R-project.org
15. Robert, C.P., Casella, G. (2004). *Monte Carlo Statistical Methods* (2nd edition). New York: Springer
16. Stern, J.M. (2008) *Cognitive Constructivism and the Epistemic Significance of Sharp Statistical Hypotheses.* Tutorial book for MaxEnt 2008, The 28th International Workshop on Bayesian Inference and Maximum Entropy Methods in Science and Engineering. July 6-11 of 2008, Boracéia, São Paulo, Brazil.