

# Scaled-Dirichlet Covariate Models for Compositional Data

## *Scaled-Dirichlet Covariate Models per Dati Compositionali*

G.S. Monti, G. Mateu-Figueras, V. Pawlowsky-Glahn, and J.J. Egozcue

**Abstract** The Dirichlet distribution provides a tool for modelling data restricted to the unit simplex  $\mathcal{S}^D$ . It has been used in several disciplines, such as geology, biology, and chemistry. The drawbacks of this distribution are extensively described in Aitchison (1986). They refer to its almost completely negative correlation and strong independence structure. A lot of generalizations were proposed to compensate for this. Among others Monti et al. (2011) proposed a Scaled-Dirichlet distribution as a perturbed random composition with a Dirichlet density. Following the approach suggested by Campbell and Mosimann (1987), and also studied by Hijazi and Jernigan (2009), here we propose a new family of regression models based on the Scaled-Dirichlet distribution. Given a vector of real covariates,  $\mathbf{s}$ , one considers the response compositional vector,  $\mathbf{X} = (X_1, \dots, X_D) \in \mathcal{S}^D$ , with a conditional Scaled-Dirichlet distribution with parameters  $\alpha(\mathbf{s}) = (\alpha_1(\mathbf{s}), \dots, \alpha_D(\mathbf{s})) \in \mathbb{R}_+^D$  and  $\beta \in \mathcal{S}^D$ . This model is characterized by the fact that its parameters are linear on  $\mathbf{s}$ , i.e. they change with the covariates. Parameter estimation and prediction issues are explored. An example with real data completes the contribution.

**Abstract** *La distribuzione di Dirichlet é storicamente nota come un modello per dati composizionali. Svariate sono le discipline in cui tale modello trova applicazione: dalla geologia, alla biologia, alla chimica. I punti di criticità di questo modello, descritti estensivamente in Aitchison (1986), sono legati alla struttura di correlazione negativa tra le sue componenti e alla forte indipendenza composizionale. In letteratura sono emerse diverse generalizzazioni al fine di compensare tale mancanza di flessibilità. Tra le altre, Monti et al. (2011) hanno proposto la*

---

G.S. Monti

Department of Economics, Management and Statistics, University of Milan-Bicocca, Italy, e-mail: gianna.monti@unimib.it

G. Mateu-Figueras and V. Pawlowsky-Glahn

Department of Computer Science, Applied Mathematics, and Statistics, University of Girona, Spain.

J.J. Egozcue

Department of Applied Mathematics III, Technical University of Catalonia, Spain.

distribuzione *Scaled-Dirichlet*, ottenuta come perturbazione di una legge di *Dirichlet*. Seguendo l'approccio proposto da Campbell and Mosimann (1987), e altresí studiato da Hijazi and Jernigan (2009), si propone una nuova famiglia di modelli basati sulla legge *Scaled-Dirichlet*. Dato un vettore di covariate,  $\mathbf{s}$ , il vettore della variabile risposta,  $\mathbf{X} = (X_1, \dots, X_D) \in \mathcal{S}^D$ , ha distribuzione condizionata di tipo *Scaled-Dirichlet* con parametri  $\alpha(\mathbf{s}) = (\alpha_1(\mathbf{s}), \dots, \alpha_D(\mathbf{s})) \in \mathbb{R}_+^D$  e  $\beta \in \mathcal{S}^D$ . Questo modello é caratterizzato dal fatto che i suoi parametri sono lineari in  $\mathbf{s}$ . Si affronta il problema di stima dei parametri e di previsione della variabile risposta. Un esempio di applicazione con dati reali completa il contributo.

**Key words:** Aitchison Geometry, Multivariate Regression, Simplex

## 1 Introduction

Compositional data are vectors of constrained, non-negative data. An appropriate representation of their sample space is the unit simplex, denoted

$$\mathcal{S}^D = \{\mathbf{x} = (x_1, \dots, x_D), x_i > 0, \sum_{i=1}^D x_i = \kappa\},$$

where  $\kappa$  is a constant, usually taken to be one or one hundred. The simplex  $(\mathcal{S}^D, \oplus, \odot, \langle \cdot, \cdot \rangle_a)$  has a  $(D-1)$ -dimensional real Euclidean vector space structure, where  $\oplus$  and  $\odot$  are the perturbation and the powering operations and  $\langle \cdot, \cdot \rangle_a$  represents the inner product. The geometry on the simplex is called *simplicial* or *Aitchison geometry* (Pawlowsky-Glahn and Egozcue, 2001). For  $\mathbf{x}, \mathbf{x}^* \in \mathcal{S}^D$ , perturbation is defined as  $\mathbf{x} \oplus \mathbf{x}^* = \mathcal{C}(x_1 x_1^*, \dots, x_D x_D^*)$ , where  $\mathcal{C}(\cdot)$  denotes the closure operation. Several disciplines involve compositional data, like e.g. rock compositions in the geological sciences, household budgets in economics, electorate compositions in political sciences, blood compositions in medicine. In the literature, also regression models which relate a compositional response variable with a system of covariates have emerged, like e.g. the Dirichlet covariate model, which assumes that the response variable follows a Dirichlet distribution whose parameters are a log-linear function of some covariates (Campbell and Mosimann, 1987; Hijazi and Jernigan, 2009; Gueorguieva et al., 2008; Melo et al., 2009). When the response variable is univariate, expressed as percentage, proportion, fraction or rate, the Dirichlet regression is called Beta regression (Ferrari and Cribari-Neto, 2004). Other options involve log-ratio transformations of the response variable using normality assumptions (Aitchison and Shen, 1980; Aitchison, 1986; Egozcue et al., 2012), or models based on the generalized Liouville distribution (Rayens and Srinivasan, 1994).

The structure of this paper is as follows: in Section 2 we revise the *Scaled-Dirichlet* distribution ( $\mathcal{SD}$ ) and in Section 3 we describe the proposed *Scaled-Dirichlet Covariate Models* for modelling the relationship between a compositional

response variable and one or more explanatory real variables. Section 4 contains an illustration of the model with a real data example.

## 2 Scaled-Dirichlet distribution

A random vector  $\mathbf{X} \in \mathcal{S}^D$  has a Scaled-Dirichlet distribution with parameters  $\alpha = (\alpha_1, \dots, \alpha_D) \in \mathbb{R}_+^D$  and  $\beta = (\beta_1, \dots, \beta_D) \in \mathcal{S}^D$  if its density function is

$$f(\mathbf{x}) = \frac{\Gamma(\alpha_+) \sqrt{D} \prod_{i=1}^D (\frac{x_i}{\beta_i})^{\alpha_i}}{\prod_{i=1}^D \Gamma(\alpha_i) (\sum_{i=1}^D \frac{x_i}{\beta_i})^{\alpha_+}}, \quad (1)$$

where  $\alpha_+ = \sum_{i=1}^D \alpha_i$ , and  $\Gamma$  denotes the gamma function. The density (1) is expressed with respect to the Aitchison probability measure (Pawlowsky-Glahn, 2003). See Monti et al. (2011) for a large discussion about the reasons and implications to use the Aitchison measure. This distribution will be denoted  $\mathbf{X} \sim \mathcal{SD}^D(\alpha, \beta)$ .

The number of parameters of the model is  $2D - 1$ . The Scaled-Dirichlet distribution can be obtained by normalizing a vector of  $D$  independent, scaled, gamma r.v.s  $W_i \sim Ga(\alpha_i, \beta_i)$ ,  $i = 1, 2, \dots, D$ ; i.e. if  $\mathbf{X} = \mathcal{C}(\mathbf{W})$ , with  $\mathbf{W} \in \mathbb{R}_+^D$ , then  $\mathbf{X} \sim \mathcal{SD}^D(\alpha, \beta)$  (Monti et al., 2011). The Scaled-Dirichlet distribution can be obtained as a perturbed random composition with a Dirichlet density, indicated by  $\mathcal{D}$ . Let  $\tilde{\mathbf{X}} \sim \mathcal{D}^D(\alpha)$  be a random composition defined in  $\mathcal{S}^D$ , and let  $\beta \in \mathcal{S}^D$  be a composition. The random composition  $\mathbf{X} = \beta \oplus \tilde{\mathbf{X}}$  has distribution  $\mathcal{SD}^D(\alpha, \beta)$ . Thus, taking into account the algebraic-geometric structure of the simplex, the Scaled-Dirichlet density is just a translation of a Dirichlet density in the simplex. The expected value of  $\mathbf{X} \sim \mathcal{SD}^D(\alpha, \beta)$  with respect to the Aitchison measure is

$$E_a(\mathbf{X}) = \beta \oplus E_a(\tilde{\mathbf{X}}), \quad (2)$$

where  $E_a(\tilde{\mathbf{X}})$  is the expected value of a Dirichlet composition:

$$E_a(\tilde{\mathbf{X}}) = \mathcal{C}(e^{\psi(\alpha_1)}, \dots, e^{\psi(\alpha_D)}). \quad (3)$$

The metric variance of  $\mathbf{X}$  coincides with the metric variance of a Dirichlet composition, because this measure of dispersion is invariant under perturbation:

$$\text{Mvar}(\mathbf{X}) = \frac{D-1}{D} (\psi'(\alpha_1), \dots, \psi'(\alpha_D)), \quad (4)$$

where  $\psi'$  is the trigamma function.

### 3 Scaled-Dirichlet Covariate Models

Following the approach of Campbell and Mosimann (1987) and Hijazi and Jernigan (2009), in order to incorporate in the model the covariate effects, we can re-parameterize each  $\alpha_i$  in terms of covariates and regression coefficients via the following log-linear model:

$$\alpha_{ij} = \alpha_{ij}(\mathbf{s}_j) = \exp\left\{\sum_{m=0}^p \delta_{im}s_j^m\right\}, \quad (5)$$

where  $\mathbf{s}_j$  is the covariate vector recorded on the  $j$ -th observed composition ( $j = 1, \dots, n$ ) and  $\delta_{im}$  are the coefficients for the  $m$ -th covariate. The parameter  $\delta_{im}$  theoretically could vary by component and also the covariates may or may not be the same set of explanatory variables. We augment each vector  $\mathbf{s}_j$  with 1 as first position for notation simplicity. So that, given a sample of size  $n$  of compositional observations distributed identically and independently, i.e.  $\mathbf{x}_1, \dots, \mathbf{x}_j, \dots, \mathbf{x}_n$ , the log-likelihood function for the re-parameterized Scaled-Dirichlet, given the covariates  $\mathbf{s}$  and ignoring the part that does not involve the parameters, is proportional to

$$\begin{aligned} l(\beta, \delta | \mathbf{x}, \mathbf{s}) \propto & \sum_{j=1}^n \left\{ \log \Gamma\left(\sum_{i=1}^D \exp\left\{\sum_{m=0}^p \delta_{im}s_j^m\right\}\right) - \sum_{i=1}^D \log \Gamma\left(\exp\left\{\sum_{m=0}^p \delta_{im}s_j^m\right\}\right) \right. \\ & - \sum_{i=1}^D \exp\left\{\sum_{m=0}^p \delta_{im}s_j^m\right\} \log \beta_i + \sum_{i=1}^D \exp\left\{\sum_{m=0}^p \delta_{im}s_j^m\right\} \log x_{ij} \\ & \left. - \left(\sum_{i=1}^D \exp\left\{\sum_{m=0}^p \delta_{im}s_j^m\right\}\right) \log\left(\sum_{i=1}^D \frac{x_i}{\beta_i}\right) \right\}, \end{aligned} \quad (6)$$

Equation (6) could be estimated using the maximum likelihood method via some optimization algorithms, e.g. the Newton-Raphson algorithm. The choice of the starting value for the algorithm is of fundamental importance to get fast convergence. For the Dirichlet regression Hijazi and Jernigan (2009) proposed a method based on re-sampling from the original data; for each re-sample a Dirichlet model with constant parameters is fitted and the mean of the corresponding covariates is computed. After that,  $D$  models of the form  $\sum_{m=0}^p \delta_{im}s_j^m$  are fitted by least squares. The fitted coefficients  $\hat{\delta}_{im}$  are used as starting values. In the case of the Scaled-Dirichlet covariate model, suitable starting values for the regression coefficients, model diagnostic and variable selection are topics open for further study.

### 4 An illustrative example

In this section we apply the Scaled-Dirichlet covariate model to analyze sedimentary data. The data originally appeared in Coakley and Rust (1968) and were presented

in Aitchison (1986). They record (sand, silt, clay) compositions at different water depths at 39 locations within an Arctic lake. The aim of this study is to model the effect of the depth covariate on the sediment composition. To solve the maximization problem defined in Eq. (6) we used the BFGS algorithm implemented in R in the *optim* routine (R Core Team, 2014). To compute the parameters standard errors, the Hessian resulting from the optimization process was used. As starting point  $\beta = \mathcal{C}(1, \dots, 1)$  was fixed, and for  $\delta$  the estimates were obtained fitting a simple Dirichlet regression:

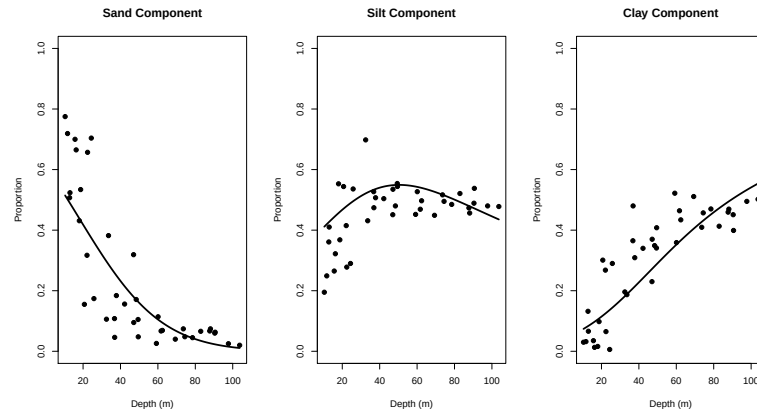
$$\log(\alpha_{ij}(\mathbf{s}_j)) = \delta_{i0} + \delta_{i1}s_j \quad i = 1, 2, 3.$$

The estimated parameters and their standard error are given in Table 1 and the fitted models (see Equation 2) are shown in Figure 1. For space constraints we do

**Table 1** Maximum likelihood estimates of Scaled-Dirichlet covariate model

| Parameter     | Estimate | S.E.  |
|---------------|----------|-------|
| $\delta_{10}$ | -0.601   | 0.369 |
| $\delta_{11}$ | -0.009   | 0.007 |
| $\delta_{20}$ | -1.005   | 0.365 |
| $\delta_{21}$ | 0.004    | 0.006 |
| $\delta_{30}$ | -1.474   | 0.341 |
| $\delta_{31}$ | 0.011    | 0.006 |
| $\beta_1$     | 0.257    | 0.232 |
| $\beta_2$     | 0.428    | 0.357 |

not report the residual diagnostic plots and the comparison with other regression models, such as the Aitchison log-ratio approach or the normal model (Egozcue



**Fig. 1** Estimated conditional means of the three components for the Arctic Lake example.

et al., 2012); but the Scaled-Dirichlet covariate model looks promising and could be an alternative to existing models. It is a natural generalization of the Dirichlet Covariate model studied by Hijazi and Jernigan (2009), that takes into account the particular algebraic-geometric structure of the simplex.

**Acknowledgements** Research partially financially supported by the Italian Ministry of University and Research, FAR (Fondi di Ateneo per la Ricerca) 2011 and by the Spanish Ministry of Economy and Competitiveness (project METRICS; Ref. MTM2012-33236).

## References

- Aitchison, J. (1986). *The Statistical Analysis of Compositional Data*. Monographs on Statistics and Applied Probability. London, Chapman & Hall Ltd. (Reprinted in 2003 with additional material by The Blackburn Press). 416 p.
- Aitchison, J. and S. M. Shen (1980). Logistic-normal distributions. Some properties and uses. *Biometrika* 67(2), 261–272.
- Campbell, G. and J. Mosimann (1987). Multivariate methods for proportional shape. pp. 10–17. ASA Proceedings of the Section on Statistical Graphics.
- Egozcue, J. J., J. Daunis-i-Estadella, V. Pawlowsky-Glahn, K. Hron, and P. Filzmoser (2012). Simplicial regression. The normal model. *J App Prob Stat* 6(1–2), 87–108.
- Ferrari, S. and F. Cribari-Neto (2004). Beta regression for modelling rates and proportions. *J App Stat* 31(7), 799–815.
- Gueorguieva, R., R. Rosenheck, and D. Zelterman (2008). Dirichlet component regression and its applications to psychiatric data. *Comput Stat Data Anal* 52(12), 5344 – 5355.
- Hijazi, R. H. and R. Jernigan (2009). Modelling Compositional Data Using Dirichlet Regression Models. *J App Prob Stat* 4, 77–91.
- Melo, T. F., K. L. Vasconcellos, and A. J. Lemonte (2009). Some restriction tests in a new class of regression models for proportions. *Comput Stat Data Anal* 53(12), 3972 – 3979.
- Monti, G. S., G. Mateu-Figueras, and V. Pawlowsky-Glahn (2011). *Notes on the Scaled Dirichlet Distribution*, pp. 128–138. John Wiley & Sons, Ltd.
- Pawlowsky-Glahn, V. (2003). Statistical modelling on coordinates. In S. Thió-Henestrosa and J. A. Martín-Fernández (Eds.), *Compositional Data Analysis Workshop – CoDaWork’03, Proceedings*. University of Girona, Girona (Spain).
- Pawlowsky-Glahn, V. and J. J. Egozcue (2001). Geometric approach to statistical analysis on the simplex. *Stoch Environ Res Risk Assess* 15(5), 384–398.
- R Core Team (2014). *R: A Language and Environment for Statistical Computing*. Vienna, Austria: R Foundation for Statistical Computing.
- Rayens, W. S. and C. Srinivasan (1994). Dependence properties of generalized liouville distributions on the simplex. *J Am Stat Ass* 89(428), pp. 1465–1470.