A SPECIAL DIRICHLET MIXTURE MODEL FOR MULTIVARIATE BOUNDED RESPONSES

Agnese Maria Di Brisco [1] and Sonia Migliorati [1]

[1] Department of Economics, Management and Statistics, University of Milano-Bicocca, (e-mail: agnese.dibrisco@unimib.it, sonia.migliorati@unimib.it)

ABSTRACT: The purpose of this paper is to provide a new regression model for multivariate continuous variables with bounded support, by taking into consideration the flexible Dirichlet, which is a special mixture of Dirichlet distributions.

KEYWORDS: bounded response, proportion, mixture, MCMC

1 Introduction

When modeling continuous variables restricted to the interval $(0,1)$, such as rates or proportions, a recent branch of research favors staying in the simplex, namely the space whose elements are unit-sum constrained vectors with strictly positive components. For the univariate case, Ferrari & Cribari-Neto, 2004, introduced a regression model for a Beta-distributed response, and recently Migliorati *et al.*, 2017a, proposed a special beta mixture regression model which provides great flexibility and a good fit in presence of outliers and in case of heavy-tailed responses. For the multivariate case, a first attempt to define a Dirichlet regression model dates back to Campbell & Mosimann, 1987. Later, Hijazi & Jernigan, 2009, extended such a model and Maier, 2014, illustrated an R package that implements a GLM-based approach to it.

Here, we propose a new multivariate regression model based on the flexible Dirichlet (FD) distribution (see Ongaro & Migliorati, 2013, and Migliorati *et al.*, 2017b), which is a special mixture of Dirichlet containing the latter as an inner point. The greater flexibility and richer parametrization of the FD over the Dirichlet lead to a promising model, as it emerges from simulations and real data applications.

2 The Flexible Dirichlet Distribution

First, let us briefly recall that the Dirichlet distributed vector $\{Y_1,\ldots,Y_D\} \sim \mathcal{D}(\boldsymbol{\alpha})$, ($0 \leq Y_j \leq 1$ for all $j = 1,\ldots,D$ and $\sum_{j=1}^{D} Y_j = 1$) has density function

(df) equal to:

$$f_D(y_1,\dots,y_D;\alpha_1,\dots,\alpha_D) = \frac{\prod_{j=1}^{D}\Gamma(\alpha_j)}{\Gamma(\alpha^+)}\prod_{j=1}^{D} y_j^{\alpha_j-1}$$

where $\alpha_1,\dots,\alpha_D > 0$ and $\alpha^+ = \sum_{j=1}^{D}\alpha_j$. An alternative parametrization, useful for regression purposes, is the mean-precision-based one, i.e.:

$$\begin{cases} \mu_j = \mathbb{E}(Y_j) = \frac{\alpha_j}{\phi} \quad j = 1,\dots,D-1 \\ \mu_D = \mathbb{E}(Y_D) = 1 - \sum_{j=1}^{D-1}\mu_j \\ \phi = \sum_{j=1}^{D}\alpha_j \end{cases} \tag{1}$$

with $0 < \mu_j < 1$, $\sum_{j=1}^{D}\mu_j = 1$ and $\phi > 0$. Here each variable is marginally beta distributed: $Y_j \sim Beta(\mu_j\phi, (1-\mu_j)\phi)$ with the first two moments equal to:

$$\mathbb{E}(Y_j) = \mu_j, \qquad Var(Y_j) = \frac{\mu_j(1-\mu_j)}{\phi+1}$$

which clearly justify why ϕ can be interpreted as a precision parameter (Ongaro & Migliorati, 2013). The reparametrized df of the Dirichlet can be written as:

$$f_D^*(y_1,\dots,y_D;\mu_1,\dots,\mu_D,\phi) = \frac{\prod_{j=1}^{D}\Gamma(\mu_j\phi)}{\Gamma(\phi)}\prod_{j=1}^{D} y_j^{\mu_j\phi-1} \tag{2}$$

Let us now consider a FD-distributed vector $\{Y_1,\dots,Y_D\} \sim \mathcal{FD}(\boldsymbol{\alpha},\mathbf{p},\tau)$ (Ongaro & Migliorati, 2013; Migliorati *et al.*, 2017b) with df

$$f_{FD}(\mathbf{y};\boldsymbol{\alpha},\mathbf{p},\tau) = \frac{\Gamma(\alpha^+ + \tau)}{\prod_{j=1}^{D}\Gamma(\alpha_j)}\prod_{j=1}^{D} y_j^{\alpha_j-1}\sum_{i=1}^{D} p_i\frac{\Gamma(\alpha_i)}{\Gamma(\alpha_i+\tau)}y_i^{\tau}$$

where $\alpha_j > 0$, $0 \le p_j < 1$ ($j = 1,\dots,D$), $\sum_j p_j = 1$, $\tau > 0$, and $\alpha^+ = \sum_{j=1}^{D}\alpha_j$. The FD distribution can be conveniently written as a finite mixture of Dirichlet distributions:

$$f_{FD}(\mathbf{y};\boldsymbol{\alpha},\mathbf{p},\tau) = \sum_{h=1}^{D} p_h f_D(\mathbf{y};\boldsymbol{\alpha}_h) \tag{3}$$

where $\boldsymbol{\alpha}_h = \boldsymbol{\alpha} + \tau\mathbf{e}_h$, and $\mathbf{e}_h$ is the canonical vector of 0's except for the $h-$th element which is 1. The special mixture structure of the FD distribution ensures that each component is distinguishable, avoiding the label switching problem. Furthermore, the FD distribution is identifiable in a strong sense and, under weak conditions, has a bounded likelihood, unlike general mixtures (see Migliorati *et al.*, 2017b, for details and proofs.)

3 The Flexible Dirichlet regression model

In order to define a FD regression (FDR) model, a mean-precision reparameterization is required. First, we reparameterize each component of the Dirichlet mixture (3) according to the alternative parametrization (1). Having defined $\phi = \alpha_1 + \cdots + \alpha_D + \tau$ and $w = \frac{\tau}{\phi}$, each component has mean vector equal to $\boldsymbol{\lambda}_h = \frac{\boldsymbol{\alpha}}{\phi} + w\mathbf{e}_h$ and precision parameter equal to ϕ. Under this reparameterization, the FD distribution can be described as a mixture of reparameterized Dirichlet:

$$f^*_{FD}(\mathbf{y};\boldsymbol{\lambda}_1,\ldots,\boldsymbol{\lambda}_D,\phi,\mathbf{p}) = \sum_{h=1}^{D} p_h f^*_D(\mathbf{y};\boldsymbol{\lambda}_h,\phi) \tag{4}$$

Furthermore, the mean vector of the FD is obtained thanks to its mixture representation (4):

$$\boldsymbol{\mu} = \frac{\boldsymbol{\alpha}}{\phi} + w\mathbf{p} \tag{5}$$

This entails a better understanding of the special mixture structure of the FD distribution. The Dirichlet-distributed components have a common precision parameter and different means such that the i-th element of the mean vector of the i-th component is higher than the corresponding element of the other mean vectors. Though, note that under this new parametrization, a constraint exists between μ_j, w and p_j. Therefore, being $0 < \frac{\alpha_j}{\alpha^+} < 1$, we get $0 < \frac{\mu_j - p_j w}{1-w} < 1$. Since μ_j will be modeled as a function of the covariates, it should be free to assume values in (0,1). So the constraint can be referred to w by normalizing it such that:

$$w^* = \frac{w}{\min_j\left\{\min\left\{\frac{\boldsymbol{\mu}}{\mathbf{p}}, \frac{1-\boldsymbol{\mu}}{1-\mathbf{p}}\right\}\right\}}$$

Under this reparameterization μ_j, p_j and w^* are free to move on $(0,1)$ while $\phi > 0$, so that the new parametric space is variation independent. Such aspect will prove to be useful both in terms of (Bayesian) estimation issues and modeling flexibility.

Let us now consider a vector of n independent multivariate responses $\mathbf{Y}^T = (\mathbf{Y}_1,\ldots,\mathbf{Y}_i,\ldots,\mathbf{Y}_n)$ such that each $\mathbf{Y}_i^T = (Y_{i1},\ldots,Y_{iD})$, for $i = 1,\ldots,n$, is a composition in the simplex. The FD regression model under the parametrization $(\boldsymbol{\mu}_i,\mathbf{p},w,\phi)$ has to take into account the constraint $\sum_{j=1}^{D}\mu_{ij} = 1$, for $j = 1,\ldots,D$. In this regard, we may adopt a multinomial logit strategy (Agresti & Hitchcock, 2005) by estimating the first $j = 1,\ldots,D-1$ parameters, having

fixed D as the baseline category:

$$\mu_{ij} = \frac{\exp\{\mathbf{x}_i^T \boldsymbol{\beta}_j\}}{\sum_{j=1}^{D} \exp\{\mathbf{x}_i^T \boldsymbol{\beta}_j\}} \quad j = 1, \ldots, D \tag{6}$$

with $\boldsymbol{\beta}_D = \mathbf{0}$. By substituting (6) into each

$$\lambda_{ijh} = \mu_{ij} - p_j w_{ij} + w_{ij} e_{jh}, \quad h = 1, \ldots, D \tag{7}$$

where $w_{ij} = w^* \min_j \left\{ \min \left\{ \frac{\boldsymbol{\mu}_{ij}}{\mathbf{p}}, \frac{1-\boldsymbol{\mu}_{ij}}{1-\mathbf{p}} \right\} \right\}$, we observe that each element of the mean vector of the h-th term of the mixture is a piecewise increasing linear functions of the μ_{ij}, varying in $(0,1)$.

The estimation problem has no explicit solution, thus a Bayesian approach to inference seems an appropriate choice. Since the FDR model is a mixture model, the allocation of data to each mixture component is unknown resulting in an incomplete data problem, for which data augmentation and Gibbs sampling are well suited. In this regard, we implemented the Gibbs sampling algorithm through the BUGS software in order to generate a finite set of values from the posterior distribution, and we further analysed the results through the R software. We iterated the algorithm until convergence by burning-in the first B simulated values, to avoid the influence of the chain's initial values. Furthermore, to properly treat autocorrelations, we also set a thinning interval, say L, such that only the first generated values in every batch of L iterations were kept. To verify the convergence of the algorithm, we checked for several statistical tests (Geweke and Heidel diagnostics for stationarity and Raftery diagnostic for autocorrelation, to name a few). To elicit the prior distribution, we assumed a priori independence, which is feasible in our context since the parametric space is variation independent and, therefore, the joint prior distribution can be factorized. We decided to adopt flat priors, so as to generate the minimum impact on the posteriors (see e.g. Albert, 2009). With respect to the regression parameters, we selected the usual multivariate normal prior with a diagonal covariance matrix with "large" values for the variances. Furthermore, we chose a gamma distribution for ϕ, a non-informative uniform prior for w and a non-informative Dirichlet for $\mathbf{p} \sim Dir(\mathbf{1})$.

4 Illustrative Application

To illustrate the FDR model we provide an application concerning $n = 39$ sediments (sand, silt and clay) collected in an Arctic lake (Aitchison, 1986). We

investigate how the composition depends on the water depth by estimating a regression model as in (6). We implemented the Dirichlet regression model and the FDR model for the vector of means in both cases by simulating an MCMC chain of length 30000 and discarding the first half iterations.

	$\boldsymbol{\beta}_0$	$\boldsymbol{\beta}_1$	$\mathbf{p}$	ϕ	w
FD	(1.704, 1.149)	(-5.215, -1.381)	(0.632, 0.055, 0.312)	34.304	0.534
Dir	(1.689, 1.258)	(-5.241, -1.588)		13.322	

Table 1. *Estimates of the parameters of the FDR model and of the Dirichlet regression model for the mean.*

From Figure 1 we observe that the FDR model and the Dirichlet one perform similarly in terms of estimation of the mean vector. However, the flexibility of the FD induces a better adaptation also to outliers observations, as it emerges from the cluster means λ_h (7).

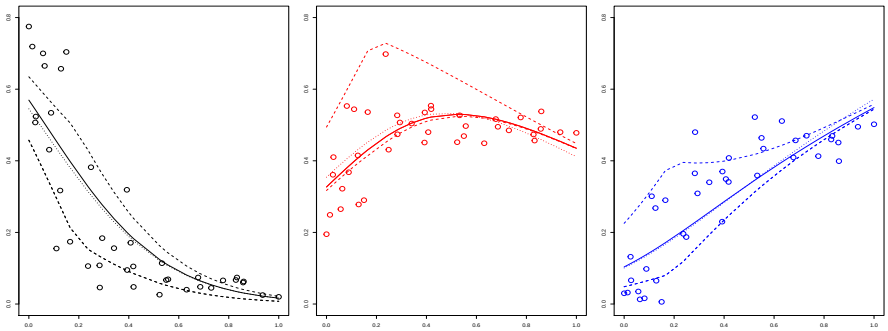

Figure 1. *Scatterplots of water depth vs sand (black), silt (red) and clay (blue). Regression lines of the μ_j for the Dirichlet (dotted lines) and for the FD regression model (solid lines). In dashed the regression lines $\boldsymbol{\lambda}_h$ for the FDR model.*

The well-known bayesian comparison criteria, namely DIC, EAIC and EBIC (Spiegelhalter *et al.*, 2002), confirm the better fit of the FDR model with respect to the Dirichlet regression model (see Table 2).

	DIC	EAIC	EBIC
FD	-218.1094	-204.4153	-189.4433
Dir	-145	-140.2061	-131.8883

Table 2. *Comparison Criteria.*

References

AGRESTI, ALAN, & HITCHCOCK, DAVID B. 2005. Bayesian inference for categorical data analysis. *Statistical Methods and Applications*, **14**(3), 297–330.

AITCHISON, JOHN. 1986. *The statistical analysis of compositional data*. Chapman and Hall London.

ALBERT, JIM. 2009. *Bayesian computation with R*. Springer Science & Business Media.

CAMPBELL, GREGORY, & MOSIMANN, JAMES E. 1987. Multivariate analysis of size and shape: modelling with the Dirichlet distribution. *Pages 93–101 of: ASA Proceedings of Section on Statistical Graphics*.

FERRARI, SILVIA, & CRIBARI-NETO, FRANCISCO. 2004. Beta regression for modelling rates and proportions. *Journal of Applied Statistics*, **31**(7), 799–815.

HIJAZI, RAFIQ H, & JERNIGAN, ROBERT W. 2009. Modelling compositional data using Dirichlet regression models. *Journal of Applied Probability & Statistics*, **4**(1), 77–91.

MAIER, MARCO J. 2014. DirichletReg: Dirichlet regression for compositional data in R. *Research Report Series/Department of Statistics and Mathematics*.

MIGLIORATI, SONIA, DI BRISCO, AGNESE MARIA, & ONGARO, ANDREA. 2017a. The Flexible Beta Regression Model. *Accepted for oral presentation at the 17th Conference of the Applied Stochastic Models and Data Analysis (ASMDA, London)*.

MIGLIORATI, SONIA, ONGARO, ANDREA, & MONTI, GIANNA S. 2017b. A structured Dirichlet mixture model for compositional data: inferential and applicative issues. *Statistics and Computing*, **27**(4), 963–983.

ONGARO, ANDREA, & MIGLIORATI, SONIA. 2013. A generalization of the Dirichlet distribution. *Journal of Multivariate Analysis*, **114**, 412–426.

SPIEGELHALTER, DAVID J, BEST, NICOLA G, CARLIN, BRADLEY P, & VAN DER LINDE, ANGELIKA. 2002. Bayesian measures of model complexity and fit (with discussion). *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, **64**(4), 583–639.