Bayesian Analysis (2024) **19**, Number 2, pp. 377–405

A Multivariate Mixture Regression Model for Constrained Responses*

Roberto Ascari†, Agnese Maria Di Brisco‡, Sonia Migliorati†, and Andrea Ongaro†

Abstract. Compositional data are vectors typically representing proportions of a whole, that is, those whose elements are strictly positive and subject to a unit-sum constraint. The increasing number of fields where this type of data arises makes the development of proper statistical tools an important issue. From a regression perspective, whenever the multivariate response is a compositional vector, a proper model that accounts for the unit-sum constraint is the well-established Dirichlet regression model. However, there are significant drawbacks mainly due to the limited flexibility of the Dirichlet distribution. The aim of this contribution is to introduce a new multivariate regression model for constrained responses, that is based on the extended flexible Dirichlet distribution (which is a structured mixture with Dirichlet distributed components). The new model is obtained by adopting a novel reparameterization which allows for, among other things, the presence of suitably designed cluster-specific regression patterns. It is shown to provide considerably greater flexibility and better performance than the standard Dirichlet regression model. In particular, from theoretical analysis, intensive simulation studies in many challenging scenarios, as well as from a real data application, it emerges that the new regression model can handle several issues affecting the Dirichlet regression, such as the presence of outliers, latent groups, multi-modality, and positive correlations.

Keywords: Dirichlet regression, simplex, outliers, latent clusters, Hamiltonian Monte Carlo.

1 Introduction

Compositional data are vectors whose elements are strictly positive and subject to a unit-sum constraint. Typically they are proportions of some whole and are encountered in several fields of science—for example, in medicine, economics, psychology, and environmetrics. Being defined on the D-part simplex $\mathcal{S}^D = \{\mathbf{y} : y_j > 0, j = 1, \ldots, D, \sum_{j=1}^{D} y_j = 1\}$, their analysis is challenging and requires proper statistical tools. This is particularly relevant in a regression context, where the constrained components of the D-dimensional response variable are regressed onto covariates, and the standard linear model is clearly unsuitable.

*This research was partially financially supported by University of Milano-Bicocca, fund number: 2019-ATE. The financial support of University of Piemonte Orientale is acknowledged by the second author.

†Department of Economics, Management and Statistics (DEMS), University of Milano-Bicocca, roberto.ascari@unimib.it; sonia.migliorati@unimib.it; andrea.ongaro@unimib.it

‡Department of Studies in Economics and Business (DISEI), University of Piemonte Orientale, agnese.dibrisco@uniupo.it

 https://doi.org/10.1214/22-BA1359

Regression on the simplex is often carried out via log-ratio data transformation, mapping $\mathcal{S}^D$ onto $\mathbb{R}^{D-1}$, so that traditional multivariate techniques can be applied (Aitchison, 2003). However, this approach can have some serious limitations as the resulting parameter estimates are only interpretable in the transformed space, whereas the main aim of the analysis is to assess and interpret the covariates' effect on the relative contributions of the original components of the whole. Moreover, the common presence of skewness, heteroscedasticity, non-normality of transformed data as well as outliers (Maier, 2014) can represent further relevant issues within the transformation approach.

An alternative solution for the regression of compositional responses is the Dirichlet model (Campbell and Mosimann, 1987; Hijazi and Jernigan, 2009), which assumes that the response variable follows a Dirichlet distribution. Typically, in accordance with a GLM-like strategy, a multinomial logit function linking the response mean vector to the covariates entails easiness of interpretation of the regression coefficients in terms of log-odds ratios. This model has been successfully applied both via maximum likelihood (Gueorguieva et al., 2008; Hijazi and Jernigan, 2009; Maier, 2014; Ankam and Bouguila, 2019; Morais et al., 2018) and within a Bayesian framework (Camargo et al., 2012; van der Merwe, 2019; Da-Silva and Rodrigues, 2015). Moreover, some extensions to account for zero patterns have been proposed as well (Tsagris and Stewart, 2018). The Dirichlet distribution is the most well-known and widespread distribution on the simplex, showing many desirable mathematical and statistical properties. However, it is unsuitable for modelling most compositional data due to its implied many forms of simplicial independences (such as compositional invariance, subcompositional invariance, partition independence, neutrality, and subcompositional independence) and limited flexibility. This is related to the fact that, once the mean vector has been fixed, a single parameter controls the whole covariance matrix, which is always characterized by negative covariances. Moreover, the Dirichlet model fails to model a wide range of relevant phenomena, including heavy tailed and multi-modal responses.

Recently, a new generalization of the Dirichlet—the extended flexible Dirichlet (EFD) distribution (Ongaro et al., 2020)—has been proposed. The EFD is a structured mixture with Dirichlet components. In a general Dirichlet mixture model each component has its own parameters which are totally unrelated to the parameters of the other components. On the contrary, in the EFD the parameters of each component are strictly linked to the remaining component-specific parameters (see Section 2.2). This structured mixture model has the advantage of displaying a clear pattern of relations among the components of the mixture, and of making the model identifiable, unlike the general Dirichlet mixture model.

The EFD distribution displays considerable flexibility in modeling dependence, as well as independence notions appropriate for compositional data, allowing also for heavy tails, multi-modality, and positive correlations. In particular, (even high) positive correlations can be reached by choosing parameter values so that the component mean vectors are arranged along a line with positive slope (for a deeper insight see Section 3.4 of Ongaro et al. (2020)). The EFD maintains several probabilistic and compositional properties of the Dirichlet, including identifiability, explicit expressions of joint

moments, as well as closure under some inferentially important operations. This leads to a substantially greater ability in capturing various data patterns, while keeping the distribution sufficiently tractable from an inferential perspective in terms of computational compliance, stability of the estimation procedure, and accuracy in evaluating estimator performances.

The aim of the present paper is to provide a new and more flexible regression model for multivariate compositional responses based on the EFD distribution. First, we propose a new parameterization of this distribution specifically designed for the regression context, which allows for a clear and meaningful interpretation of parameters. Then, we define a regression model based on it, which substantially extends the Dirichlet model both because of the greater variety of possible shapes of the EFD, and the unique built-in regression behavior. Indeed, on the one hand we directly model the response mean vector as an appropriate function (i.e., the multinomial logit) of covariates, thus ensuring interpretable effects of the covariates on the response variables. On the other hand, we allow for the presence of suitably-designed, cluster-specific regression patterns that are capable of capturing possible deviations from the main trend (for example, outliers or latent groups). Finally, the model further enriches the analysis, if deemed convenient, by regressing onto covariates also other important aspects such as the precision parameter and/or relevant characteristics of the clusters.

We adopt a Bayesian approach for inference. In particular, a novel feature of the model is a general probabilistic scheme defining prior distributions for the component mixing weights, which lets the number and type of cluster-specific regression components to be freely selected within the model.

The potential of the new model is illustrated by means of extensive simulation studies as well as a real data application with various choices of covariates. The new model proves to be superior to the Dirichlet model in all examined settings. Indeed, it is capable of recognizing, by suitably arranging its mixture components, various data patterns, such as latent groups, outliers, multi-modality, and positive correlations. Instead, typically these data patterns are not captured by the Dirichlet model. In any case, the new model often displays substantially better fit and higher precision in parameter estimates compared to the Dirichlet model.

The rest of this paper is organized as follows. In Section 2, we review the Dirichlet and EFD distributions, proposing a regression-devised parameterization for the latter. In Section 3, we introduce the new regression model and a variant which has the objective of smoothing the cluster-specific regression curves. An identifiability property is proved as well. Section 4 briefly points to some possible further uses of the EFD in the context of discrete compositional data. Section 5 is concerned with the Bayesian estimation procedure and the details about the prior elicitation scheme with special regard to cluster selection. Three intensive simulation studies characterized by different purposes are presented and discussed in Section 6. Section 7 presents an application to a compositional dataset studied in the ecological field. Finally, concluding remarks are provided in Section 8, while the Supplementary Material (SM) (Ascari et al., 2023) file provides further details especially (but not only) on simulations and applications together with some proofs. Tables and figures in SM are labeled with prefix "S".

2 Distributions on the simplex

In this section we briefly review the Dirichlet and EFD distributions, and propose a reparameterization of the latter, expressly devised for regression purposes.

2.1 Dirichlet and EFD distributions

A Dirichlet distributed D-dimensional random vector $\mathbf{Y} \sim \text{Dir}(\cdot; \boldsymbol{\alpha})$ has the following probability density function (p.d.f.):

$$f_D(\mathbf{y}; \boldsymbol{\alpha}) = \frac{\Gamma(\alpha^+)}{\prod_{j=1}^D \Gamma(\alpha_j)} \prod_{j=1}^D y_j^{\alpha_j - 1}, \tag{1}$$

where $\mathbf{y} \in \mathcal{S}^D$, $\boldsymbol{\alpha} = (\alpha_1, \ldots, \alpha_D)^\intercal$, $\alpha_j > 0$, and $\alpha^+ = \sum_{j=1}^D \alpha_j$. The Dirichlet distribution can freely model the mean vector, which is $\mathbb{E}[Y_j] = \frac{\alpha_j}{\alpha^+}$, but it dedicates only the parameter α^+ to handle the entire variance-covariance matrix.

The EFD distribution (Ongaro et al., 2020) can be viewed as a particular Dirichlet mixture. Specifically,

$$\text{EFD}(\mathbf{y}; \boldsymbol{\alpha}, \boldsymbol{\tau}, \mathbf{p}) = \sum_{r=1}^D p_r \text{Dir}(\mathbf{y}; \boldsymbol{\alpha} + \tau_r \mathbf{e}_r), \tag{2}$$

where $\mathbf{y} \in \mathcal{S}^D$, $\mathbf{e}_r$ is a vector of zeros except for the r-th element which is equal to 1, $\mathbf{p} = (p_1, \ldots, p_D)^\intercal$ is such that $0 \leq p_r < 1$ and $\sum_{r=1}^D p_r = 1$, and the vector $\boldsymbol{\alpha} = (\alpha_1, \ldots, \alpha_D)^\intercal$ has positive elements. The constraint $p_r < 1$ (implying a minimum of two components) is imposed to avoid identifiability issues. Note that we decided to allow some p_r's to be equal to zero so that the model allows for a number of components up to a maximum of D. If p_r is null for some r (i.e., the corresponding cluster is not present), the value of the parameter τ_r is immaterial, and it needs to be fixed to avoid non identifiability. We decided to fix it to 1 for convention. Therefore, the definition of the parameter space is completed by setting $\tau_r > 0$ for any r such that $p_r > 0$, and $\tau_r = 1$ for any r such that $p_r = 0$. The p.d.f. of an EFD distribution is:

$$f_{EFD}(\mathbf{y}; \boldsymbol{\alpha}, \boldsymbol{\tau}, \mathbf{p}) = \left(\prod_{j=1}^D \frac{y_j^{\alpha_j - 1}}{\Gamma(\alpha_j)} \right) \sum_{r=1}^D p_r \frac{\Gamma(\alpha_r)\Gamma(\alpha^+ + \tau_r)}{\Gamma(\alpha_r + \tau_r)} y_r^{\tau_r}. \tag{3}$$

The mixture (2) defining the EFD is "structured," in the sense that it entails suitable links among the mixture components (note that hereafter "component," also referred to as "cluster" or "group," will be reserved to indicate an element of a mixture). The links are expressed by the parameter vector $\boldsymbol{\alpha}$, which is common to all components, and can be interpreted (after normalization) as a barycenter of the model. The specificity of each component is given by the parameters τ_r and the vectors $\mathbf{e}_r$. The vector $\mathbf{e}_r$ conveys the information that the r-th cluster is different from the others in its r-th element, which

displays a higher mean (with respect to the r-th mean element of the other clusters). The extent to which this mean is higher than the others is dictated by the parameter τ_r. This fact will be discussed in detail in Section 2.2, where a new parametrization of the EFD will be proposed (Equations (4) and (5)).

The EFD, besides displaying a far richer dependence structure than the Dirichlet, greatly extends the variety of shapes of the Dirichlet p.d.f., particularly in terms of multi-modality, asymmetry, and heavy tails. Concerning multi-modality, recall that the Dirichlet is unimodal if and only if all α_r's are larger than one. It follows that, by suitably choosing the parameters (possibly, but not necessarily, setting some p_r's to zero), any number of modes up to D can be reached. As for the tails' behavior, the Dirichlet presents some limitations. To analyze this behavior, we study (1) on the frontier of the simplex, when one or more elements go to zero. In this case, any unimodal Dirichlet has tails always tending to zero. On the contrary, in the EFD, by choosing some α_r's equal to one and others larger than one, it is possible to reach unimodal densities which can be proved to have strictly positive (and finite) tails corresponding to the elements with unitary α_r. A figure illustrating two such interesting cases is reported in Section 1 of the SM.

Finally, the EFD distribution shows several theoretical properties, such as some simplicial forms of dependence/independence, simple expressions of marginal, conditional, and sub-compositional distributions, and identifiability. Concerning identifiability, the EFD can be shown to be identifiable in the following strong sense: Two EFD distributions are the same if and only if the corresponding parameters are equal. Thus, for example, no labeling issues occur as there is no invariance under permutation of the components. This is generally not true for general mixture models, and can be proved to fail specifically in case of arbitrary Dirichlet mixtures. All these properties make the EFD tractable from computational and inferential viewpoints. For more details see Ongaro et al. (2020).

It is worth noting that the EFD distribution contains the Dirichlet as an inner point when $\tau_r = 1$ and $p_r = \alpha_r/\alpha^+$ for every $r = 1, \ldots, D$. Moreover, the flexible Dirichlet (FD) distribution (Ongaro and Migliorati, 2013) is obtained by setting all the τ_r's equal (i.e., $\tau_1 = \cdots = \tau_D = \tau$). This latter distribution, being a structured mixture itself, displays several theoretical properties of interest for compositional data (Migliorati et al., 2017). However, its cluster structure has been shown to be too restrictive for modeling many types of data, as noted in Ongaro et al. (2020). In addition, the FD, like the Dirichlet, only admits negative covariances. This led us to focus on the more general EFD model, which, among other things, allows for high positive correlations.

2.2 EFD reparametrization

With the aim of regressing a compositional vector onto a set of covariates, it is convenient to work with parameterizations based on mean vectors. In the Dirichlet case, we thus prefer to adopt the mean-precision parametrization, namely $\bar{\boldsymbol{\alpha}} = (\bar{\alpha}_1, \ldots, \bar{\alpha}_D)^\intercal \in \mathcal{S}^D$, where $\bar{\alpha}_j = \mathbb{E}[Y_j] = \frac{\alpha_j}{\alpha^+} > 0$ for $j = 1, \ldots, D$, and $\alpha^+ = \sum_{j=1}^D \alpha_j > 0$, the latter representing the precision parameter. We denote the mean-precision parametrized Dirichlet with $\text{Dir}_{mp}(\cdot; \bar{\boldsymbol{\alpha}}, \alpha^+)$. With this parametrization, the EFD can be rewritten as

$$\text{EFD}(\mathbf{y}; \boldsymbol{\alpha}, \boldsymbol{\tau}, \mathbf{p}) = \sum_{r=1}^{D} p_r \text{Dir}_{mp}(\mathbf{y}; \boldsymbol{\lambda}_r, \alpha^+ + \tau_r), \tag{4}$$

where

$$\boldsymbol{\lambda}_r = (1 - w_r)\bar{\boldsymbol{\alpha}} + w_r \mathbf{e}_r, \tag{5}$$

and $w_r = \frac{\tau_r}{\alpha^+ + \tau_r}$. Note that $\boldsymbol{\lambda}_r = (\lambda_{r1}, \ldots, \lambda_{rD})^\intercal$, which represents the mean of the r-th mixture component, can be interpreted as a weighted average of a common barycenter, $\bar{\boldsymbol{\alpha}}$, and the r-th simplex vertex $\mathbf{e}_r$. Formula (5) clearly highlights the links and differences among the component means: Each $\boldsymbol{\lambda}_r$ departs from the common barycenter $\bar{\boldsymbol{\alpha}}$ in its r-th element, which is increased by a factor controlled by w_r, being $\lambda_{rr} = \bar{\alpha}_r + w_r(1 - \bar{\alpha}_r)$. The exceedence $w_r(1 - \bar{\alpha}_r)$ is distributed among the other elements proportionally to the $\bar{\alpha}_h$'s, that is, the ratios between the other elements of $\boldsymbol{\lambda}_r$ are the same as the ratios of the corresponding $\bar{\alpha}_h$ $(h = 1, \ldots, D; h \neq r)$. Clearly, the larger w_r, the farther apart $\boldsymbol{\lambda}_r$ is from the common barycenter and from all the other component means.

For regression purposes, it is important to understand the consequences of this mixture structure on the relation among the group means for any given element j of the composition. For the symmetry properties of the model, we can assume, without loss of generality, that $\tau_1 \geq \ldots \geq \tau_D$ so that $w_1 \geq \ldots \geq w_D$. Then, given a generic element j, one can easily see that $\lambda_{jj} > \bar{\alpha}_j > \lambda_{Dj} \geq \ldots \geq \lambda_{(j+1)j} \geq \lambda_{(j-1)j} \geq \ldots \geq \lambda_{1j}$, $(j = 1, \ldots, D)$. Thus, always with reference to the j-th element, the mean of the j-th group is higher than all the others. Furthermore, the ordering among the other group means is dictated only by the w_r's. Indeed, it holds that $\frac{\lambda_{rj}}{\lambda_{hj}} = \frac{1 - w_r}{1 - w_h}$ $(j = 1, \ldots, D;\ r, h \neq j)$. Obviously, if $w_r = w_h$, then the corresponding component means collapse for all elements different from $\{r, h\}$, thus reducing the number of different component means for a given element. In the extreme case when all the w_r's are equal (i.e., in the FD model), we just have two different component means for each j-th element, namely the one relative to the j-th group and all the others which coincide.

Since our main objective is to regress the mean vector of the model onto covariates, we need to find a further suitable parameterization of the EFD explicitly including the mean vector $\boldsymbol{\mu}$. The first order moment of the EFD easily follows from its mixture structure and Dirichlet distribution properties:

$$\mu_j = \mathbb{E}[Y_j] = \sum_{r=1}^{D} p_r \lambda_{rj} = \bar{\alpha}_j \sum_{r=1}^{D} p_r(1 - w_r) + p_j w_j, \quad j = 1, \ldots, D. \tag{6}$$

However, the parameterization of the EFD based on μ_j, p_j and w_j ($j = 1, \ldots, D$) is not variation independent, which may generate inferential difficulties (especially within a Bayesian MCMC-based approach to inference), and may also prevent separate modeling of any parameter as a function of covariates. This is due to the following constraints:

$$0 < \bar{\alpha}_j = \frac{\mu_j - p_j w_j}{1 - \sum_r p_r w_r} < 1 \qquad j = 1, \ldots, D, \tag{7}$$

whereas the constraint $\sum_{j=1}^{D} \bar{\alpha}_j = 1$ is automatically satisfied. A careful investigation of (7) shows that, to ensure that these inequalities are satisfied, it is sufficient to ask that $\bar{\alpha}_j > 0$, thus $p_j w_j < \mu_j$. This implies that a variation independent parameter space can be obtained by reparameterizing either the w_j's or the p_j's. If $\boldsymbol{\mu}$ is fixed, meaning that is it not regressed onto covariates, either choices are equivalent. Instead, when $\boldsymbol{\mu}$ depends on covariates, different regression models are obtained by reparameterizing either the w_j's or the p_j's. We decide to keep the p_j's fixed and let the w_j's vary together with the μ_j's. The reasons for this choice are better discussed in Section 3. A natural reparameterization of the w_j's is given by

$$\tilde{w}_j = \frac{w_j}{m(\mu_j)}, \qquad j = 1, \ldots, D, \tag{8}$$

where $m(\mu_j) = \min\left\{\frac{\mu_j}{p_j}, 1\right\}$, so that $\tilde{w}_j \in (0, 1)$ represents the extent to which each mixture component departs from the common barycenter $\bar{\boldsymbol{\alpha}}$.

The final parameterization of the EFD$(\cdot; \boldsymbol{\mu}, \mathbf{p}, \alpha^+, \tilde{\mathbf{w}})$ is therefore based on $\boldsymbol{\mu} \in \mathcal{S}^D$, $\mathbf{p} = (p_1, \ldots, p_D)^\intercal$ such that $0 \leq p_r < 1$ and $\sum_{r=1}^{D} p_r = 1$, $\alpha^+ > 0$, and $\tilde{\mathbf{w}} = (\tilde{w}_1, \ldots, \tilde{w}_D)^\intercal$ with $\tilde{w}_j \in (0, 1)$, where we conventionally set $\tilde{w}_j = 1/2$ if $p_j = 0$.

3 The EFD regression model

A widely-used regression model for compositional responses is the Dirichlet regression (DirReg) model that has two different ways of implementation. One way regresses the parameters α_j's of the standard Dirichlet parameterization (based on Equation (1)) onto covariates (Hijazi and Jernigan, 2009; Gueorguieva et al., 2008; Camargo et al., 2012; Maier, 2014; Ankam and Bouguila, 2019; Morais et al., 2018). The other way regresses the Dirichlet mean values α_j/α^+'s onto covariates (van der Merwe, 2019; Tsagris and Stewart, 2018; Da-Silva and Rodrigues, 2015; Maier, 2014). In our opinion, the latter is preferable as it provides more interpretable results in agreement with the general GLM strategy (McCullagh and Nelder, 1989). Indeed, the α_j's do not have a clear and relevant meaning. As suggested by a referee, two other generalizations of the Dirichlet distribution (Graf, 2020; Ankam and Bouguila, 2019) have been used in the literature to construct a regression model. Though, neither of them regresses the mean value of original data onto covariates (see Section 9 of the SM for more details on these two models and a comparison with the EFDReg model on the real data application studied in Section 7).

Our objective is to define a generalization of the mean-based Dirichlet regression model by considering the EFD distribution for the response variables. Specifically, let

$\mathbf{Y} = (\mathbf{Y}_1, \ldots, \mathbf{Y}_n)^\intercal$ be the response matrix such that $\mathbf{Y}_i = (Y_{i1}, \ldots, Y_{ij}, \ldots, Y_{iD})^\intercal$ is a D-dimensional vector on the simplex ($i = 1, \ldots, n$), and let $\mathbf{X} = (\mathbf{x}_1, \ldots, \mathbf{x}_n)^\intercal$ be the design matrix, where $\mathbf{x}_i = (1, x_{i1}, \ldots, x_{it}, \ldots, x_{iT})^\intercal$ is a $(T+1)$ dimensional vector of covariates values on the i-th sample unit. The mean vector $\boldsymbol{\mu}_i$ of $\mathbf{Y}_i$ can be regressed onto covariates by choosing a smooth and invertible link function, which transforms the expectation of the response variable to the linear predictor. Therefore, in our context an adequate mapping from the simplex to the $D-1$ Euclidean space must be selected. Although any such function is compatible with our methodology, in the following we make use of the multinomial logit function –say g– as it provides a simple interpretation of the regression coefficients in terms of log-odds ratios. We then have

$$g(\mu_{ij}) = \log\left(\frac{\mu_{ij}}{\mu_{iD}}\right) = \mathbf{x}_i^\intercal \boldsymbol{\beta}_j, \tag{9}$$

where $\mu_{ij} = \mathbb{E}[Y_{ij}]$ and $\boldsymbol{\beta}_j = (\beta_{j0}, \beta_{j1}, \ldots, \beta_{jT})^\intercal$ is a vector of regression coefficients ($j = 1, \ldots, D-1$). Thus, the expectation of the response can be written as:

$$\mu_{ij} = g^{-1}(\mathbf{x}_i^\intercal \boldsymbol{\beta}_j) = \begin{cases} \frac{\exp(\mathbf{x}_i^\intercal \boldsymbol{\beta}_j)}{1+\sum_{r=1}^{D-1}\exp(\mathbf{x}_i^\intercal \boldsymbol{\beta}_r)}, & j = 1, \ldots, D-1 \\ \frac{1}{1+\sum_{r=1}^{D-1}\exp(\mathbf{x}_i^\intercal \boldsymbol{\beta}_r)}, & j = D. \end{cases} \tag{10}$$

Note that the D-th element is conventionally fixed as baseline. Indeed, formally any element of the composition can be placed in the D-th position as the EFD model is symmetric in its elements. Thus, if there exists an element which is more important or with a particularly significant meaning, then this element should be used as reference for ease of interpretation. A further point to take into consideration when choosing the reference element stems from possible computational instability due to low values of the reference element. To reduce this problem a simple preprocessing step consists in choosing as reference element the one with the highest sample mean.

The EFD regression (EFDReg) model is then defined by assuming that each $\mathbf{Y}_i$ is independently distributed as an EFD$(\cdot; \boldsymbol{\mu}_i, \mathbf{p}, \alpha^+, \tilde{\mathbf{w}})$. With this notation, the DirReg model is obtained instead by letting $\mathbf{Y}_i$ follow a Dirichlet distribution $\text{Dir}_{mp}(\cdot; \boldsymbol{\mu}_i, \alpha^+)$, that is, with mean $\boldsymbol{\mu}_i$ and precision parameter α^+.

In a regression perspective, the general mean $\boldsymbol{\mu}$ varies with covariates. As a consequence of the new parametrization of the EFD including $\boldsymbol{\mu}$, the component means $\boldsymbol{\lambda}_r$ also vary with covariates ($r = 1, \ldots, D$) following the relationships studied in Section 2.2 and further explored below. Our model can thus be viewed as a type of mixture regression model (Frühwirth-Schnatter, 2006), with specific relations occurring among group regressions. However, our focus is different from general mixture regression models. Indeed, in the latter models the main interest is on separately modeling the component regressions which display arbitrarily different dependencies on covariates. The main inferential objective of our model is to assess the general impact of covariates' on the mean vector of the response variable. In this respect, a mixture of regression models usually does not produce clearly interpretable results.

Though our main regression focus is on the overall mean, interesting and interpretable implications of our model in terms of component-specific regressions can be

derived, as described in Section 2.2, with reference to the general behavior of $\boldsymbol{\lambda}_r$. The behavior of the w_r's which affect $\boldsymbol{\lambda}_r$ (see Formula (5)) are yet to be examined. Indeed, the w_r's are not fixed, since they depend on $\boldsymbol{\mu}$ in the adopted mean-based parameterization, particularly $w_r = \tilde{w}_r\, m(\mu_r)$. This choice of w_r has the advantage of reaching the whole parameter space by varying $\tilde{w}_r$. However, it induces a non-smooth piecewise linear dependence of $\boldsymbol{\lambda}_r$ on μ_r and, therefore, on the covariates, which may not be always desirable. Thus, the identification of a further alternative parameterization giving rise to smooth component-specific regression curves seems of interest. This can be accomplished by choosing w_r proportional to the linear or parabolic function which best approximates the upper constraint for w_r given by $m(\mu_r)$. The maximum linear function below $m(\mu_r)$ is the identity $w_r = \mu_r$. However, this solution seems too restrictive. We therefore turned to parabolic curves, deriving the uniformly highest parabolic function under the requirement of being below $m(\mu_r)$, thus obtaining the following function:

$$q(\mu_r) = \begin{cases} \frac{\mu_r}{p_r}[1 - (1 - p_r)\mu_r] & \text{if} \quad p_r \geq 1/2 \\ \mu_r(2 - \mu_r) & \text{if} \quad p_r < 1/2, \end{cases} \tag{11}$$

(see Section 2 of the SM for the full derivation, as well as a graphical illustration). This parabolic choice originates a new regression model, hereafter denoted by EFDReg with parabolic constraints (EFDReg$^{\mathrm{P}}$), by assuming that the responses are independently distributed as EFD$(\cdot; \boldsymbol{\mu}_i, \mathbf{p}, \alpha^+, \dot{\mathbf{w}})$, where

$$\dot{w}_r = \frac{w_r}{q(\mu_r)}, \qquad r = 1, \ldots, D. \tag{12}$$

As shown in Figure S2, the trade-off for the smoothness of the new proposal is that it cannot cover the whole parameter space, thus resulting in a slightly less flexible model. Though note that, in our experience both in simulation studies and in real data analyses, the parabolic approximation $q(\mu_r)$ shows a behavior comparable to the piecewise linear function $m(\mu_r)$, except for the case of extreme values for w under the piecewise linear function data generating mechanism. However, we investigated the possibility of more sophisticated approximations, such as piecewise polynomial functions. The problem is not trivial. One has to find the uniformly highest piecewise polynomial, which is at the same time everywhere differentiable, nonnegative, and below the piecewise linear natural constraint. We proved that a solution exists for the piecewise parabolic case (see Section 3 of the SM, where further comments as well as a graphical illustration are reported). Thus, this solution can be adopted if one is particularly concerned with this type of flexibility, though with some additional computational burden. An analysis of the proof used to derive the piecewise parabolic function highlights that no uniformly optimal solution can be expected to exist for higher order piecewise polynomials.

To achieve a variation independent parameter space, all the above parametrizations are based on the choice of letting the w_j's vary with covariates and the p_j's be fixed, as anticipated in Section 2.2. Obviously, the constraints imposed by variation independence, namely $p_j w_j < \mu_j$, could be fulfilled by a variety of different modeling choices for p_j and w_j as functions of μ_j, the simplest other choice being to fix the w_j's and let the p_j's vary. Although this choice is certainly possible within our methodology,

and may be adequate in some specific contexts, we believe that it is generally more appropriate to fix the p_j's and we stick to this choice in the present paper. Indeed, often latent groups of fixed sizes are present, with possibly different behaviors with respect to covariates. To better exemplify the implied differences between the two frameworks, consider the simple case when the overall mean μ_j of the j-th element goes to zero. It seems reasonable to expect that, in this case, all the group means tend to zero, which is exactly what happens when the w_j's are let free, as all the w_j's then tend to zero. Conversely, by fixing the w_j's, $\mu_j \to 0$ has the implication that $p_j \to 0$, that is, one group must vanish. This seems a less common scenario. Analogous remarks apply to the case $\mu_j \to 1$.

The above EFD-based regression models only regress the mean response onto covariates. However, the variances of the responses are functions of the corresponding means, thus they are indirectly modeled as functions of covariates too. In fact, a natural form of heteroscedasticity is present, which accounts for the fact that the variance of Y_{ij} is strictly dependent on its mean μ_i; its maximum being $\mu_{ij}(1-\mu_{ij})$. However, it is possible to directly model the variance by letting α^+ depend on covariates. Indeed, α^+ is a precision parameter, as the variance can be easily shown to be a decreasing function of α^+. In particular, a regression model for α^+ can be defined as:

$$h(\alpha_i^+) = \mathbf{x}_i^\intercal \boldsymbol{\gamma}, \tag{13}$$

where $h(\cdot)$ is a strictly monotone function from $\mathbb{R}^+$ to $\mathbb{R}$, usually the logarithm, and $\boldsymbol{\gamma}$ is a vector of regression coefficients. There may also be specific contexts where it may be desirable to model the weights p_j's and even the group departures w_j's as functions of covariates. Again, this can be easily accommodated within the EFD regression model.

Finally, it is important to underline that the identifiability of the EFD distribution implies that the EFD-based regression models are identifiable under the necessary condition that the design matrix has full rank. This is a rather exceptional property for a mixture model, which greatly helps in interpreting the parameters and, simultaneously, in avoiding computational difficulties such as label-switching issues. Below, we provide an identifiability result for the EFDReg model. We assume for simplicity that only the mean depends on covariates. However, the result can be analogously proved for the EFDReg$^{\mathrm{P}}$ model and easily extended to the general case where some other parameters are modeled as functions of covariates (see Section 4 of the SM for the proof).

Proposition 3.1. *Consider a vector* $(\boldsymbol{Y}_1, \ldots, \boldsymbol{Y}_n)^\intercal$ *of independent response variables* $\boldsymbol{Y}_i \sim EFD(\boldsymbol{\mu}_i, \boldsymbol{p}, \alpha^+, \tilde{\mathbf{w}})$, $i = 1, \ldots, n$. *Suppose that*

$$\boldsymbol{\mu}_i = g(\mathbf{x}_i^\intercal \boldsymbol{\beta}_1, \ldots, \mathbf{x}_i^\intercal \boldsymbol{\beta}_{D-1}),$$

where $\boldsymbol{x}_i$ *is a covariate vector of dimension* $T+1 \leq n$, *and* $g(\cdot)$ *is an invertible function from* $\mathbb{R}^{D-1}$ *to* $\mathcal{S}^D$. *Let us denote by EFDReg*$(\boldsymbol{\eta})$, *where* $\boldsymbol{\eta} = (\boldsymbol{\beta}_1, \ldots, \boldsymbol{\beta}_{D-1}, \boldsymbol{p}, \alpha^+, \tilde{\mathbf{w}})^\intercal$, *the corresponding regression model, and let* $\mathbf{Y} \sim EFDReg(\boldsymbol{\eta})$ *and* $\mathbf{Y}' \sim EFDReg(\boldsymbol{\eta}')$. *Then, if the design matrix* $\mathbf{X}$ *is of full rank* $(T+1)$, $\mathbf{Y} \sim \mathbf{Y}'$ *if and only if* $\boldsymbol{\eta} = \boldsymbol{\eta}'$.

4 EFD and discrete compositional data

In this section we briefly point to some possible further uses of the EFD not directly related to continuous compositional data. We discuss them in this paper to better highlight the potential of the EFD model, as also suggested by a referee. The EFD may be considered as a prior distribution for the probability vector of a multinomial model, which can be viewed as a distribution for discrete compositional data, thus generalizing the usual Dirichlet prior. It can be shown that the EFD prior is still conjugate with a very simple posterior expression. Specifically, let us consider a sample $\mathbf{t}_i$ $(i = 1, \ldots, n)$ whose elements are drawn independently from a multinomial with D categories and parameters n_i and $\boldsymbol{\pi} = (\pi_1, \ldots, \pi_D)^\intercal$, denoted by $\text{Mult}(n_i, \boldsymbol{\pi})$. If $\boldsymbol{\pi}$ is given an $\text{EFD}(\boldsymbol{\alpha}, \boldsymbol{\tau}, \mathbf{p})$ prior distribution, then it can be shown that the posterior $\boldsymbol{\pi} \mid \mathbf{t}_1, \ldots, \mathbf{t}_n$ is an $\text{EFD}(\boldsymbol{\alpha} + \mathbf{t}^+, \boldsymbol{\tau}, \mathbf{p}^*/p_+^*)$, where $\mathbf{t}^+ = (t_1^+, \ldots, t_D^+)^\intercal$, t_j^+ is the total count of the j-th category,

$$p_j^* = p_j \frac{(\alpha_j + \tau_j)^{[t_j^+]}}{(\alpha^+ + \tau_j)^{[t^+]} \alpha_j^{[t_j^+]}},$$

$p_+^* = \sum_{j=1}^{D} p_j^*$, $t^+ = \sum_{j=1}^{D} t_j^+$, and $\alpha^{[y]} = \Gamma(\alpha + y)/\Gamma(\alpha)$. Thus, in the posterior the hyperparameter $\boldsymbol{\alpha}$ is updated with the same simple mechanism of the parameter of a Dirichlet prior. Data also update the mixing weights $\mathbf{p}^*/p_+^*$. In particular, each weight p_j^*/p_+^* depends increasingly on the corresponding t_j^+, given the other t_r^+, $r \neq j$ $(r, j = 1, \ldots, D)$.

Moreover, the EFD distribution may provide a relevant contribution also in the analysis of multivariate count data, that can be thought of as discrete compositions summing to a fixed integer. This can be fulfilled via a compound approach that imposes a proper simplex distribution (i.e., a prior distribution) on the parameters of a multinomial model (as discussed above), but focusing on the resulting predictive distribution to model data. A Dirichlet prior is a popular choice leading to the well-known Dirichlet-multinomial (DM) distribution. The DM is a widespread and more flexible alternative to the multinomial distribution. Specifically, if $\mathbf{T} = (T_1, \ldots, T_D)^\intercal$ follows a multinomial distribution with parameters n and $\boldsymbol{\pi} = (\pi_1, \ldots, \pi_D)^\intercal$, and $\boldsymbol{\pi} \sim \text{Dir}_{mp}(\boldsymbol{\mu}, \alpha^+)$, then the $\text{DM}(n, \boldsymbol{\mu}, \alpha^+)$ has probability mass function:

$$f_{\text{DM}}(\mathbf{t}; n, \boldsymbol{\mu}, \alpha^+) = \frac{n!\,\Gamma(\alpha^+)}{\Gamma(n + \alpha^+)} \prod_{r=1}^{D} \frac{\Gamma(t_r + \mu_r \alpha^+)}{t_r!\,\Gamma(\mu_r \alpha^+)}. \tag{14}$$

By compounding the multinomial with an EFD distribution, an EFD-multinomial distribution $\text{EFDM}(\cdot; n, \boldsymbol{\alpha}, \boldsymbol{\tau}, \mathbf{p})$ is obtained. As a consequence of the Dirichlet mixture structure in Equation (4) of the EFD, the new model can be written as an analogous structured finite mixture with DM components:

$$f_{\text{EFDM}}(\mathbf{t}; n, \boldsymbol{\alpha}, \boldsymbol{\tau}, \mathbf{p}) = \sum_{r=1}^{D} p_r f_{\text{DM}}(\mathbf{t}; n, \boldsymbol{\lambda}_r, \alpha^+ + \tau_r), \tag{15}$$

where $\boldsymbol{\lambda}_r$ is given by (5). A detailed analysis of the properties of this distribution goes beyond the scope of the paper. Here we only highlight that this model can be used as response distribution in a regression framework. To this end, it is convenient to derive the mean vector of the EFDM distribution, which will then be linked to covariates. As the mean of the $\text{DM}(\cdot; n, \boldsymbol{\mu}, \alpha^+)$ is $n\boldsymbol{\mu}$, the EFDM mean is $n\sum_{r=1}^{D} p_r \boldsymbol{\lambda}_r$, thus being proportional to the EFD mean (6). It follows that a mean-based reparametrization strategy similar to the one proposed for the EFDReg model (see Sections 2.2 and 3) can be adopted to define an EFDM regression model, keeping a similar interpretation of parameters and component specific regressions.

5 Inferential issues

To obtain estimates of the unknown parameters of the DirReg, EFDReg, and EFDReg^{P} models we favor a Bayesian approach. This choice is mainly motivated by the difficulty, both computational and analytical, of likelihood-based inferential approaches in dealing with complex models such as mixtures. Conversely, the finite mixture structure of the EFD distribution can be advantageously treated as an incomplete data problem within the Bayesian paradigm (Frühwirth-Schnatter, 2006). Monte Carlo (MC) methods are particularly suited for dealing with mixture models. Among these methods, a recent solution is the Hamiltonian Monte Carlo (HMC) algorithm (Neal, 1994), a generalization of the Metropolis algorithm which combines Markov Chain Monte Carlo (MCMC) and deterministic simulation methods (for details on HMC and its implementation see Section 7.1 of the SM).

The posterior distributions of the unknown parameters are simulated based on full likelihood and prior distributions, which lead to the following full joint distribution:

$$\prod_{i=1}^{n} f^*(\mathbf{y}_i; \boldsymbol{\eta})\ \pi(\boldsymbol{\eta}),$$

where $f^*(\cdot;\cdot)$ denotes the p.d.f. of the assumed distribution (Dirichlet or EFD), $\boldsymbol{\eta} = (\boldsymbol{\beta}_1, \ldots, \boldsymbol{\beta}_{D-1}, \mathbf{p}, \alpha^+, \tilde{\mathbf{w}})^\intercal$ is the vector of all the unknown parameters, and $\pi(\boldsymbol{\eta})$ is its prior distribution. Regarding priors' choice, we take advantage of the variation-independent parameter space and suppose prior independence, that is:

$$\pi(\boldsymbol{\eta}) = \pi(\boldsymbol{\beta}_1, \ldots, \boldsymbol{\beta}_{D-1})\pi(\mathbf{p})\pi(\alpha^+)\pi(\tilde{\mathbf{w}}).$$

Moreover, we adopt non- or weakly-informative priors to induce minimum impact on the posteriors (Albert, 2009). In particular, we select a multivariate normal with zero mean vector and diagonal covariance matrix with large values of the variances $\boldsymbol{\sigma}_j^2$ as non-informative prior for the regression parameters $\boldsymbol{\beta}_j$ $(j = 1, \ldots, D-1)$. If the precision parameter α^+ is also regressed onto a set of covariates, an equivalent non-informative multivariate normal prior (with variances $\boldsymbol{\tau}^2$) is set for the regression coefficients $\boldsymbol{\gamma}$. Otherwise, the precision parameter α^+ follows a $\text{Gamma}(g, g)$ distribution, with g small enough to induce a large variability. Moreover, we adopt a $\text{Uniform}(0, 1)$ prior for w_j^N, namely the normalized version of w_j, $j = 1, \ldots, D$, which is $\tilde{w}_j$ for the EFDReg model and $\dot{w}_j$ for the EFDReg^{P} model (see Formulas (8) and (12), respectively).

The elicitation of a prior for the weights $\mathbf{p}$ is a more critical task. Indeed, a standard choice, such as the uniform distribution corresponding to a Dirichlet with hyperparameters all equal to one, implies that, with probability one, all the D possible components of the mixture defining the EFD distribution are present. The same holds true by adopting any absolutely continuous (with respect to the Lebesgue measure) distribution on the D-dimensional simplex. On the contrary, as stressed in Section 2, we think it is an important element of flexibility to let the number K ($2 \leq K \leq D$) of the mixture components (i.e., the number of non zero weights p_j's) to be freely selected within the model. Incidentally, notice that our model can display up to D components, where D coincides with data dimensionality. Thus, we will endow K with a prior distribution, so that its value may be inferred by inspecting the corresponding posterior. In addition, since the number of modes of the distribution is a function (also) of K, this number will be random in our framework, and may be therefore inferred from data. This is relevant from a theoretical, computational, as well as applicative point of view. In particular, according to our experience, when the real number of clusters is lower than D, convergence problems in posterior simulation algorithms arise, leading to extremely dispersed posterior distributions. Therefore, we decided to devise a general probabilistic scheme generating prior distributions which randomly select the clusters to be included in the model. Incidentally, note that the constraint $K \geq 2$ has been imposed to keep identifiability of the model, as mentioned in Section 2.1. Further details are captured below.

A general way to define such a scheme is to first extract the number of clusters according to a discrete random variable K with values in $\{2, \ldots, D\}$. Here, a standard choice is the uniform distribution, but alternatives may be easily implemented if prior information is available. Then, conditionally on K, the probability of including a specific set of clusters has to be defined, that is we need to express a probability measure over the unordered sets G_K composed by K distinct elements chosen from $\{1, \ldots, D\}$. In the absence of prior information on the probabilities of inclusion of the different clusters, one can treat them symmetrically. Therefore, a non informative prior is obtained assuming a uniform distribution on G_K. Specifically, we have:

$$P\left(G_K = \{i_1, \ldots, i_K\}\right) = \frac{1}{\binom{D}{K}}, \tag{16}$$

where the i_j's are all distinct with range $\{1, \ldots, D\}$. In this case it can be proved that the probability that G_K includes the generic i-th cluster is K/D. Although, in some contexts, prior information may be available on the probability that each cluster is present in the mixture. One way to incorporate this information in the prior is to express probabilities $0 < \theta_r < 1$, with $\sum_{r=1}^{D} \theta_r = 1$, which quantify the relative propensity of each cluster to enter the mixture. We then propose a natural probabilistic scheme to sequentially sample the K clusters, which can be interpreted as follows (for a formal definition see Section 6.1 of the SM). Consider a population formed by D groups, the r-th group having frequency θ_r. Then, choose the first K distinct groups extracted from an i.i.d. sample of this population. Interestingly, if we select that the θ_r's all equal (thus, equal to $1/D$), we obtain for G_k the uniform non-informative distribution in (16). The latter is therefore a special case of the general scheme here introduced.

The prior for the weights $\mathbf{p}$ is then completed by choosing, conditionally on K and on G_K, a distribution for the selected mixing weights on the K dimensional simplex, the standard non-informative choice being the uniform distribution.

In applications we found it convenient to allow K to also take the value 1, in which case the model is set to coincide with the DirReg model. Even though this choice leads to loss of identifiability, it is computationally and inferentially highly beneficial. Indeed, by so doing we include the DirReg model as a particular case of the EFDReg model, and can give positive prior probability to it. Please note that in absence of covariates, the Dirichlet distribution is a particular case of the EFD, namely $\text{Dir}_{mp}(\cdot; \boldsymbol{\mu}, \alpha^+) = \text{EFD}(\cdot; \boldsymbol{\mu}, \mathbf{p} = \boldsymbol{\mu}, \alpha^+, \tilde{w}_r = 1/(1+\alpha^+), r = 1, \ldots, D)$ (see Section 2.1). Instead, the DirReg model is not included in the EFDReg, unless $\mathbf{p}$ is forced to depend on covariates through the same regression used to model $\boldsymbol{\mu}$. In addition, it can be shown that in this new setting, non-identifiability is restricted to those very special configurations of regression coefficients and covariate values leading to $\mu_i = \mu$ for every $i = 1, \ldots, n$, and therefore it is practically irrelevant. The above described choice will be made in the following simulations and application, together with uniform priors on K, on G_K (see Formula (16)), and on the selected p_r's.

However, this prior can face computational inefficiency for large values of D, as the number of possible clusters to be considered increases exponentially with D. Therefore, we devised a further new prior which scales well with D, although it is not as general and flexible as the original one. Essentially, we jointly select K and the set of clusters G_K by simply drawing i.i.d. Bernoulli random variables $B_i \sim \text{Bin}(1, \theta)$, $i = 1, \ldots, D$, so that cluster i is included if $B_i = 1$. As this mechanism allows for the possibility of choosing $K = 0$ (i.e., no cluster), to force the presence of at least one cluster we slightly modify it by first including a cluster chosen at random. A more precise formulation is given in Section 6.2 of the SM, where we also prove that the new prior is a particular case of the original prior, obtained by assigning to K a (shifted) binomial distribution and, conditional on K, a uniform distribution to G_K. Notice also that the parameter θ, which determines the new prior, can be chosen by simply fixing the expected (prior) number of clusters. This new prior allows convergence of algorithms even for large values (up to 100) of D. However, the involved computational cost rapidly increases with D and with the larger sample sizes required to reliably estimate the consequent growing number of parameters (for example, a value of $D = 100$ involves hundreds of parameters). A simulation study performed with $n = 500$ showed reliable results with D up to 25 (see Section 10 of the SM for details).

In conclusion, the two EFD-based regression models are summarized by the directed acyclic graph (DAG) in Figure 1. The observed variables, the random variables (including latent variables), and the hyperparameters of the prior distributions are represented as rectangles, non-filled circles, and filled circles, respectively. The $\mathbf{Z}_i$ are the latent membership variables of the mixture components. Moreover, the red dashed elements are optional and highlight the possibility of enriching the model with a regression for the precision parameter as in Equation (13). In this case, for ease of reading, we omitted to add the subscript i to α^+ and to place it inside the thick rectangle.

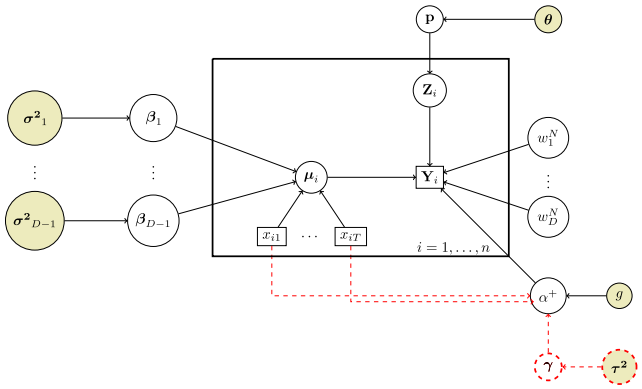

Figure 1: DAG for EFD-based regression models. Observed variables (rectangle-shaped nodes), random variables including latent variables (non-filled circle-shaped nodes), and fixed hyperparameters (filled circle-shaped nodes) are represented. Red dashed edges and red nodes represent optional elements to account for a regression model for the precision parameter α^+.

6 Simulation studies

To compare the performances of the EFDReg, EFDReg$^{\mathrm{P}}$, and DirReg models, we simulated a variety of scenarios that cover many potentially challenging problems, among which are multi-modality, presence of heavy tails, outliers, and latent groups. In what follows, we illustrate the samples' simulation schemes and the main inferential results for each scenario. We take advantage of the HMC algorithm for estimating the vector of unknown parameters $\boldsymbol{\eta}$, which varies according to the considered model. More specifically, in the DirReg model $\boldsymbol{\eta} = (\boldsymbol{\beta}_1, \ldots, \boldsymbol{\beta}_{D-1}, \alpha^+)^\intercal$, whereas in the EFDReg model $\boldsymbol{\eta} = (\boldsymbol{\beta}_1, \ldots, \boldsymbol{\beta}_{D-1}, \alpha^+, \mathbf{p}, \tilde{\mathbf{w}})^\intercal$, and in the EFDReg$^{\mathrm{P}}$ model $\boldsymbol{\eta} = (\boldsymbol{\beta}_1, \ldots, \boldsymbol{\beta}_{D-1}, \alpha^+, \mathbf{p}, \dot{\mathbf{w}})^\intercal$. For the sake of simplicity, we shall denote the normalized version of $\mathbf{w}$ by $\mathbf{w}^N$, thus having $\mathbf{w}^N = \tilde{\mathbf{w}}$ or $\mathbf{w}^N = \dot{\mathbf{w}}$ in the EFDReg and in the EFDReg$^{\mathrm{P}}$ models, respectively. Each scenario is replicated 200 times to obtain MC measures regarding the estimation performances, such as MC estimators' mean, their root mean squared errors (rMSEs), and the coverage level of the 95% credible sets (CSs).

The algorithm is implemented via the Stan modeling language (Stan Development Team, 2016). Details on the Stan implementation and on convergence diagnostics can be found in Section 7.1 of the SM. Moreover, Stan codes for the described/proposed models are available upon request to the authors.

For space constraints in the next subsections we report brief comments on all considered scenarios, whereas a deeper analysis is devoted to only a few cases. Full details and a self-contained presentation of results can be found in Sections 7.3–7.9 of the SM.

6.1 Fitting studies

First, some fitting studies are considered by simulating data from the DirReg (scenario (i)), EFDReg (scenario (ii)), EFDReg$^{\mathrm{P}}$ (scenario (iii)), and the additive logistic-normal (ALN) (Aitchison, 2003) (scenario (iv)) regression models. The objective of these studies is to analyze the goodness of fit and estimates of regression coefficients of each model under different data generating mechanisms. The sample size is set to $n = 150$, and the multivariate response lies on the 3-part simplex (so that results are easier to be visualized). In scenarios (i)–(iii), we regress the mean (see Equation (9)) onto a quantitative covariate X, uniformly distributed in $(-0.5, 0.5)$ with regression coefficients set equal to $\beta_{10} = -1$, $\beta_{11} = 1.5$, $\beta_{20} = 0.5$, and $\beta_{21} = -3$. In scenario (i), the response is Dirichlet distributed with precision parameter $\alpha^+ = 50$. In scenarios (ii) and (iii) (EFDReg and EFDReg$^{\mathrm{P}}$ respectively), the remaining parameters of the models are fixed equal to $\alpha^+ = 50$, $\mathbf{p} = (0.5, 0.3, 0.2)^\intercal$, and $\mathbf{w}^N = (0.6, 0.2, 0.9)^\intercal$. In scenario (iv), data are generated from an ALN regression model, that is

$$\left(\log\left(\frac{Y_{i1}}{Y_{i3}}\right), \log\left(\frac{Y_{i2}}{Y_{i3}}\right)\right)^\intercal \sim N_2\left(\begin{pmatrix} 1 + 2x_i \\ 0.5 - 3x_i \end{pmatrix}, \begin{bmatrix} 1.5 & 2 \\ 2 & 4 \end{bmatrix}\right). \tag{17}$$

Hereafter, the goodness of fit of each model will be quantified through the Watanabe-Akaike information criterion (WAIC) (Watanabe, 2010; Vehtari et al., 2017). Models with better fit are associated with smaller WAIC values (see Section 7.2 of the SM for more details). The WAIC values of all models in scenario (i) are almost identical, despite the data generating mechanism should favor the DirReg model. The precision of all parameter estimates is very similar among competing models as well, and the regression curves are almost completely overlapped. The posterior mode of the discrete parameter K is equal to 1 under both the EFDReg and EFDReg$^{\mathrm{P}}$ models, thus indicating that both models adapt to data without overfitting (see Section 7.3 of the SM).

Focusing on scenario (ii) (EFDReg generating model), it is worth noting the presence of clusters characterizing the response (Figure 2). It must also be underlined that the chosen parameters' configuration results in a very extreme location of one cluster, namely the third one corresponding to $w_3^N = 0.9$. The purpose of this choice is to test the EFDReg$^{\mathrm{P}}$ model's behavior in a case which is theoretically known to be most critical for such a model given its parameter space constraints. Table 1 shows that the EFDReg model provides very good estimates of all parameters (smallest biases, smallest rMSEs, and highest coverages) together with (by far) the lowest WAIC in all replications. Moreover, coherently with the data structure, the posterior distribution of K is degenerate at 3. On the contrary, the EFDReg$^{\mathrm{P}}$ model fails to fully capture the data pattern (the posterior mode of the number of mixture components is equal to 2). However, it still performs substantially better than the DirReg model, as highlighted by Table 1 and Figure 2. Note also in Figure 2 the peculiar and rather extreme behavior of the third component of the model (red dotted lines in center and right panels).

As for scenario (iii), where data are generated from the EFDReg$^{\mathrm{P}}$ model, both the EFD-based regression models use three mixture components to properly fit the data in almost all replications (see Section 7.5 of the SM). Overall, the EFDReg$^{\mathrm{P}}$ produces better fitting and estimates, but the discrepancy with respect to the EFDReg is not as

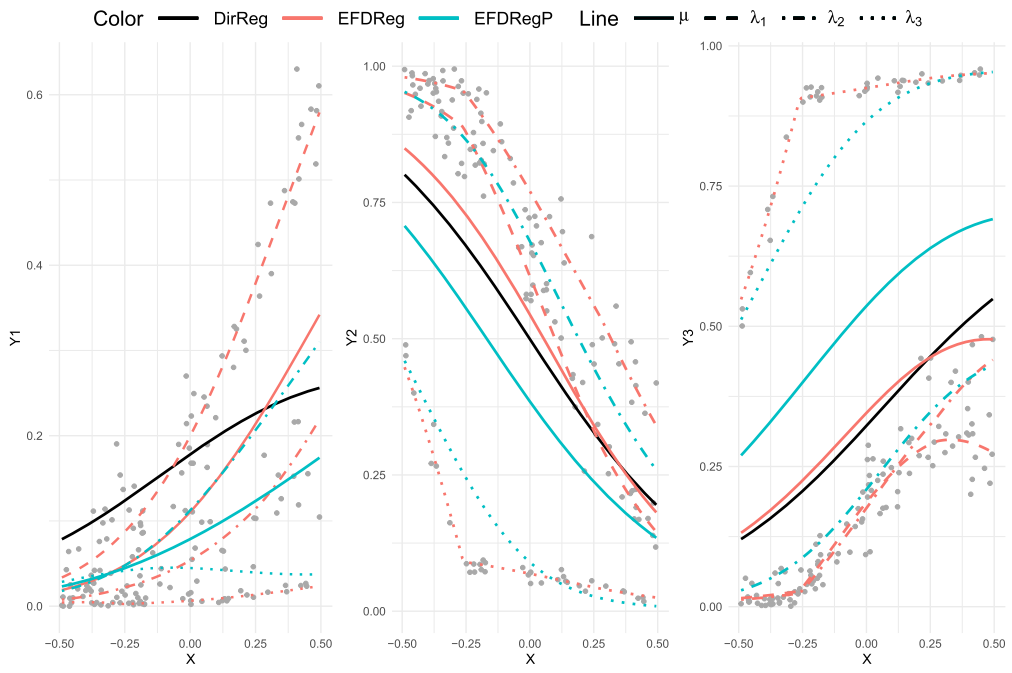

Figure 2: One randomly selected replication from the EFDReg fitting study (scenario (ii)). Estimated EFDReg (red), EFDReg$^{\mathrm{P}}$ (blue), and DirReg (black) models. Solid curves refer to $\boldsymbol{\mu}$. Regression curves for $\boldsymbol{\lambda}_1$ (dashed), $\boldsymbol{\lambda}_2$ (dot-dashed), and $\boldsymbol{\lambda}_3$ (dotted) are present only if the corresponding components are selected by the model.

	DirReg			EFDReg			EFDReg$^{\mathrm{P}}$		
Parameter	MC mean	rMSE	coverage	MC mean	rMSE	coverage	MC mean	rMSE	coverage
$\beta_{10} = -1$	-0.490	0.516	0	-1.008	0.077	0.985	-1.751	0.773	0.07
$\beta_{11} = 1.5$	-0.328	1.853	0	1.504	0.141	0.96	1.081	0.522	0.6
$\beta_{20} = 0.5$	0.562	0.126	0.8	0.503	0.071	0.935	-0.299	0.817	0.06
$\beta_{21} = -3$	-3.157	0.354	0.915	-3.023	0.11	0.95	-2.843	0.276	0.685
$\alpha^{+} = 50$	3.358	46.644	0	50.594	4.412	0.955	7.940	42.11	0
K	—			3 (100%)			2 (77%)		
Mixture comp.	—			—			2 and 3		
WAIC	-544.440 (0%)			-1127.342 (100%)			-736.678 (0%)		

Table 1: Results from the EFDReg fitting study (scenario (ii)). MC means, rMSEs, and coverage probabilities under each model are reported. For the parameter K, the posterior mode together with its posterior probability (in parenthesis) and the chosen components when the mode is less than $D = 3$ are reported. Last row shows the average WAIC and the % of times that model was the best model.

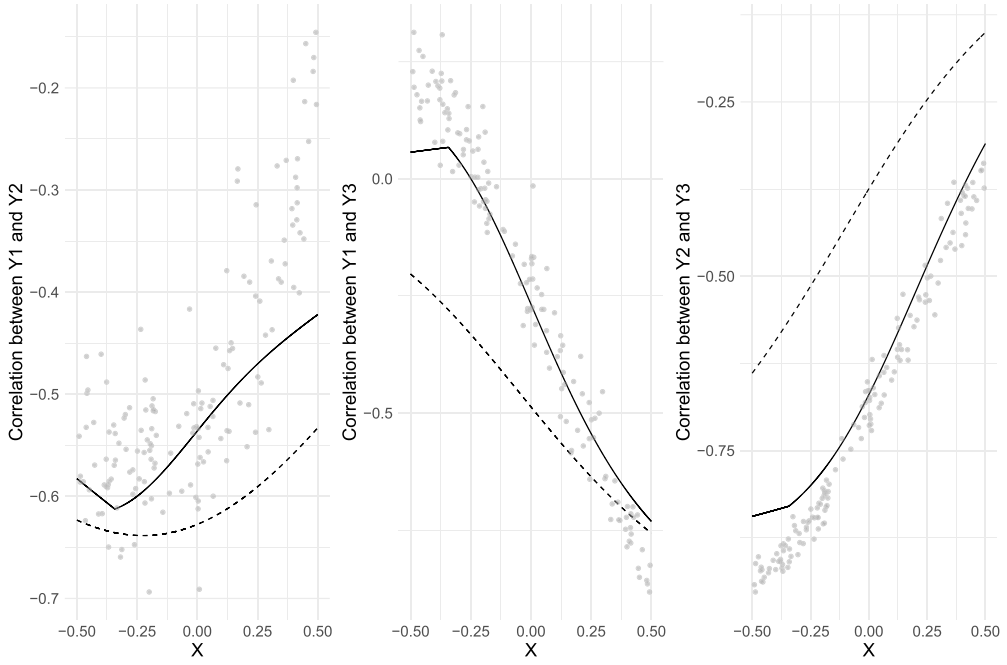

Figure 3: Correlation coefficients of the fitted DirReg (dashed lines) and EFDReg (solid lines) models as a function of the continuous covariate X in fitting scenario (iv). Each point refers to a correlation computed on a sample from the ALN model for a fixed value of X.

large as in the previous (quite extreme) scenario (ii). The DirReg produces substantially poorer results than the other two models. We highlight that in both scenarios (ii) and (iii), the DirReg model displays particular difficulties in estimating the regression coefficients (namely β_{11} and β_{21}).

In scenario (iv), all three models adequately capture the pattern of dependence between response variables and the covariate, despite the regression implied by the ALN model is quite different from the regression implied by the three Dirichlet-based models. Though, the EFDReg and EFDReg^P models show better fits and predictive ability than the DirReg model (see Section 7.6 of the SM). Additionally, the ability of the EFDReg model (the EFDReg^P model giving similar results) in recovering the pattern and the sign of the correlation coefficients between the compositional variables is assessed. Figure 3 shows the correlation coefficients of the fitted DirReg (dashed line) and EFDReg (solid line) models as the continuous covariate X varies. The points correspond to pairwise sample MC correlation coefficients relative to ALN-generated data given X. Interestingly, the EFDReg model produces a very accurate fitting unlike the DirReg model, and it succeeds in identifying positive correlations (see center panel of Figure 3).

6.2 Linear predictor misspecification and latent groups

The second simulation study aims at exploring the presence of latent structures in the data. We focus on two different cases: The first case, scenario (a), considers a latent (unobserved) covariate (linear predictor misspecification), whereas the second case, scenario (b), assumes a finite mixture of two Dirichlet components with the same mean, but different precision parameters. In scenario (a), the additional covariate has been included in the data generating process, but omitted in all three regression models. In particular, we replicated the generating mechanism of scenario (i) by adding a latent dichotomous covariate, whose categories have relative frequencies of 0.3 and 0.7. The additional regression coefficients are set equal to $\beta_{12} = 4$ and $\beta_{12} = 2.5$ (scenarios (a.1) and (a.2), respectively) and $\beta_{22} = 2$. Here we comment on scenario (a.1), but results from scenario (a.2) are consistent, as illustrated in Section 7.7 of the SM, where full details on both scenarios are presented.

The presence of latent variables in data is particularly challenging for the DirReg which, in fact, shows by far the worst adaptation to data (WAIC equal to -451.0). Conversely, the EFDReg$^{\mathrm{P}}$ model provides a slightly better fit to data (WAIC equal to -848.8) than the EFDReg one (WAIC equal to -819.0), and results show that it is the preferable model in 74.5% of replications. Indeed, the latent covariate affects the estimates of regression coefficients and precision parameter of all models (see Table S6). However, the EFDReg and EFDReg$^{\mathrm{P}}$ models recognize the presence of two latent groups, which can not be captured by the DirReg model (see Section 7.7 of the SM).

As for scenario (b), the regression coefficients for the mean vector are the same as used in scenario (i). The first DirReg component is characterized by a precision $\alpha_1^+ = 50$ and a mixing weight of 0.6, whereas the second by a smaller precision $\alpha_2^+ = 3.5$ and a mixing weight of 0.4, thus inducing heavy tails. Note that all data share the same regression curve, which is an assumption in agreement with the DirReg regression model. Despite that, all three models produce reliable estimates of the regression coefficients (see Table 2). In addition, the EFD-based regression models better capture the inflated variability through their mixture components, thus providing a better WAIC (see Section 7.8 of the SM).

6.3 Outliers

In a compositional framework, (multivariate) outliers can typically arise due to measurement errors, heavy tails, or multi-modality. To study this aspect, we generated outliers through the perturbation operator, which is the compositional version of the addition operation, and is defined as $\mathbf{y} \oplus \boldsymbol{\delta} = \mathcal{C}\left((y_1 \cdot \delta_1, \ldots, y_D \cdot \delta_D)^\intercal\right) \in \mathcal{S}^D$, where $\mathbf{y}$ and $\boldsymbol{\delta}$ are vectors on the simplex playing the roles of perturbed and perturbing elements, respectively. Moreover, the closure operation $\mathcal{C}(\cdot)$ is defined as $\mathcal{C}(\mathbf{q}) = (q_1/q^+, \ldots, q_D/q^+)^\intercal$ with $q^+ = \sum_{j=1}^D q_j$ and $q_j > 0, j = 1, \ldots, D$. The neutral element of the perturbation operation is $\boldsymbol{\delta} = (1/D, \ldots, 1/D)^\intercal$, so that if element y_j is perturbed by δ_j greater (lower) than $1/D$, the perturbation is upward (downward). In this scenario, data are generated from a DirReg model with the same $\boldsymbol{\beta}_1, \boldsymbol{\beta}_2$, and α^+ parameters as in scenario (i).

	DirReg			EFDReg			EFDReg$^{\mathrm{P}}$		
Parameter	MC mean	rMSE	coverage	MC mean	rMSE	coverage	MC mean	rMSE	coverage
$\beta_{10} = -1$	-0.994	0.1	0.8980	-0.949	0.126	0.8673	-0.950	0.118	0.8520
$\beta_{11} = 1.5$	1.469	0.377	0.8724	1.329	0.429	0.8112	1.295	0.475	0.7755
$\beta_{20} = 0.5$	0.505	0.073	0.9592	0.506	0.073	0.9694	0.496	0.077	0.9235
$\beta_{21} = -3$	-3.008	0.274	0.9490	-2.901	0.277	0.9184	-3.015	0.263	0.9082
$\alpha_1^+ = 50$	6.538			7.626			7.690		
$\alpha_2^+ = 3.5$									
K	—			3 (76.02%)			3 (79.08%)		
WAIC	-769.648 (8%)			-790.309 (38%)			-790.581 (54%)		

Table 2: Results from the latent groups study (scenario (b)). MC means, rMSEs, and coverage probabilities under each model are reported. For the parameter K, the posterior mode together with its posterior probability (in parenthesis) are reported. Last row shows the average WAIC and the % of times that model was the best model.

Then, 15 observations (i.e., 10% of the data) are perturbed according to two different schemes. In the first one (scenario (I)), we perturb observations with covariate X greater than its third tertile, whereas in the second scenario (scenario (II)), we perturb uniformly over the range of X. In both cases, we use $\boldsymbol{\delta} = (0.82, 0.09, 0.09)^\intercal$ (complete results are given in Section 7.9 of the SM).

As expected, since the chosen $\boldsymbol{\delta}$ value highly perturbs the first element of the composition upwards, all three models show some difficulties in estimating β_{10} and β_{11} regression coefficients, with the EFDReg displaying worse performances particularly for β_{11}. However, both EFD-based regression models reach a substantially better and comparable fit in terms of WAIC (lower than -800 in both cases, versus a value around -666 for the DirReg) and a higher estimate of the precision parameter α^+. This is because the EFD-based regression models can recognize two sub-populations, namely the main data body and the outlying observations. In particular, the first mixture component is dedicated to the outlying observations in coherence with the choice of $\boldsymbol{\delta}$, which perturbs upward only the first element of the composition. Moreover, the estimated value of the mixing weight p_1 takes values between 0.1 and 0.2 for both the EFD-based models under both scenarios, which is close to the true proportion of outliers.

These remarks are further confirmed by inspecting the presence of possibly influential observations through the conditional-predictive ordinate (CPO) diagnostic (Gelman et al., 2013). CPO is a Bayesian measure used to detect unlikely observations given the fitted model, which is defined as the predictive density of the i-th unit once it has been excluded from the dataset (see Section 7.2 of the SM for details). Figure 4 compares the estimated CPOs under each model for the 15 outlying observations of scenarios (I) and (II). Both EFD-based regression models show a general better fit to outliers since their estimated CPO values are almost always much greater than the DirReg's one. This also suggests that outliers are less influential under the two former models.

In summary, all considerations drawn in this section support using the EFDReg and EFDReg$^{\mathrm{P}}$ models, as they do not perform worse than the DirReg model when the latter is true, while displaying a clearly superior behavior in all other cases. The two

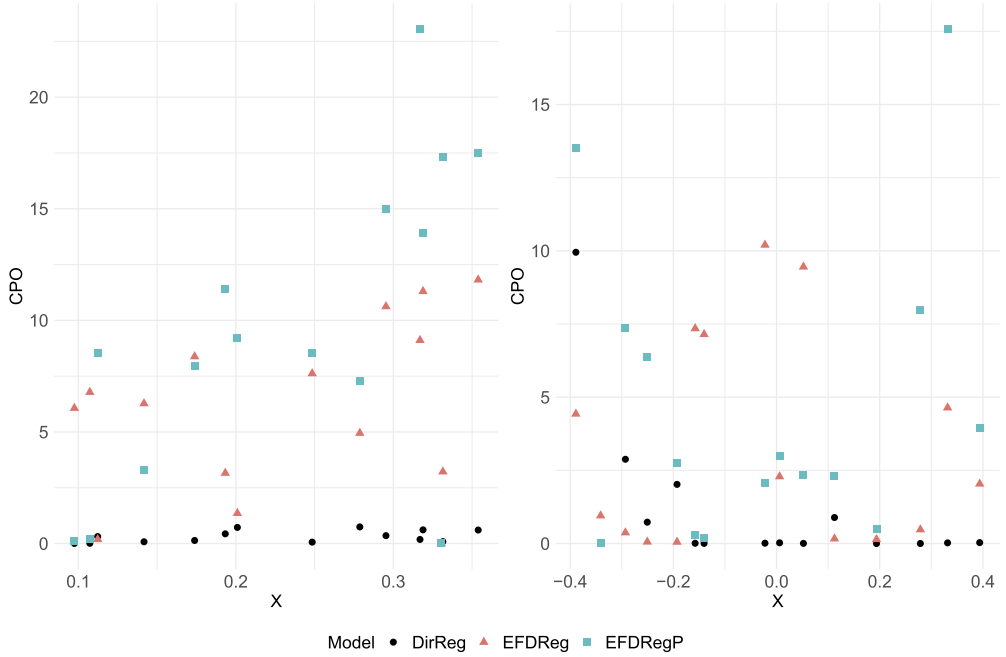

Figure 4: Distribution of the estimated CPO values by model (shapes and colors) for the outlying observations in a randomly selected replication from the outlier study (left panel: Scenario (I); right panel: Scenario (II)). X-axis reports the corresponding values of the covariate.

new EFD-based regression models show comparable performances except in extreme cases (scenario (ii) of the fitting study in Section 6.1). We thus generally prefer the EFDReg$^{\mathrm{P}}$ model due to its smooth regression functions, even though, in practice, both models can be estimated and compared in any single application.

7 Application to plants data

We further studied the performance of our models using a real data application where the proportion of biomass in roots (RMF), stems (SMF), and leaves (LMF) were evaluated on a sample of $n = 500$ plants (Poorter et al., 1995). Douma et al. (2019) used this 3-part composition to illustrate the DirReg model. The goal of the original study was to detect differences in the composition between slow- and fast-growing species, while considering different nitrate supply levels (high or low). For this reason, Poorter et al. considered two different species of plants: *Deschampsia flexuosa* (*D. flexuosa*) and *Holcus lanatus* (*H. lanatus*), that are slow- and fast-growing species, respectively. Replicated plants were harvested at different times (after 21–49 days) and biomass was recorded, including the total amount of biomass (TDM). The latter is the size of the composition, which is not

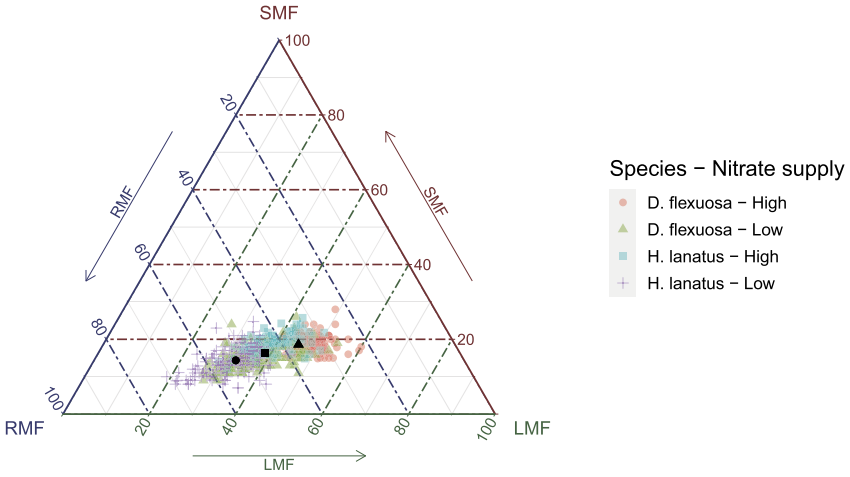

Figure 5: Ternary diagram of plants data. Black square refers to predicted mean under EFD models, whereas black circle and triangle represent EFDs $\boldsymbol{\lambda}_1$ and $\boldsymbol{\lambda}_3$, respectively.

constant for every plant. Figure 5 shows the ternary diagram of the data, with shapes and colors representing different combinations of species and nitrate supply. We can see that almost every plant has a mild-to-low amount of biomass attributable to its stems (10–30%), whereas the remaining biomass composition depends on the specific combination of species and nitrate supply level. For example, *H. lanatus* plants treated with low nitrate supply are characterized by a larger proportion of roots biomass, while *D. flexuosa* plants with high nitrate supply are characterized by a higher proportion of leaves biomass.

We first estimated the parameters without considering any covariates (scenario A). Note that in this case the two EFD models are equivalent apart from a slightly smaller parameter space for the EFD with parabolic constraints, thus we only discuss the EFD model. From Table 3, we can see that the EFD model detects two latent groups of similar size ($p_1 \approx 0.54$ and $p_3 \approx 0.46$). Plotting the estimated $\boldsymbol{\lambda}_1$ and $\boldsymbol{\lambda}_3$ in the ternary diagram in Figure 5, we can describe these groups as the *H. lanatus* plants treated with low nitrate supply and the *D. flexuosa* plants treated with high nitrate supply, respectively. This shows how the EFD can properly detect sub-populations even without explanatory variables, leading to a far better WAIC than the Dirichlet. In addition, it is interesting to analyze how the two models capture correlations among the three biomass proportions (see Table 4 where sampling and estimated correlations are reported). The Dirichlet model produces substantially worse estimates than the EFD. In particular, and different from the EFD, it completely fails to recognize the positive association between LMF and SMF, and it highly overestimates the two remaining (negative) correlations. Finally, we graphically compare the fitting of uni- and bi-dimensional densities (Figures S21–S23). The ability of the EFD model to properly capture bi-modality when present is apparent, but a substantial better fit than the Dirichlet is provided in the unimodal

	DirReg		EFDReg	
Parameter	Mean	95% CS	Mean	95% CS
μ_1	0.448	(0.441, 0.455)	0.452	(0.444, 0.460)
μ_2	0.168	(0.163, 0.173)	0.163	(0.160, 0.167)
μ_3	0.384	(0.378, 0.391)	0.385	(0.378, 0.391)
α^+	42.24	(38.605, 45.936)	91.217	(82.744, 99.973)
p_1	—	—	0.541	(0.489, 0.593)
p_2	—	—	—	—
p_3	—	—	0.459	(0.407, 0.511)
w_1^N	—	—	0.298	(0.269, 0.327)
w_2^N	—	—	—	—
w_3^N	—	—	0.032	(0.006, 0.059)
K	—		2 (89.35%)	
Mixture components	—		1 and 3 (87.98%)	
WAIC	−2729.4		−3096.3	

Table 3: Plants data, no covariate case (scenario A). Posterior means and 95% CSs of the parameters under the DirReg and EFDReg models. For the parameter K the posterior mode, its posterior probability and the chosen components are reported.

	LMF vs SMF	LMF vs RMF	SMF vs RMF
Sample	0.497	−0.949	−0.745
Dirichlet	−0.355 (−0.361, −0.348)	−0.711(−0.719, −0.705)	−0.405 (−0.411, −0.398)
EFD	0.215 (0.153, 0.269)	−0.905 (−0.916, −0.893)	−0.601 (−0.651, −0.560)

Table 4: Plants data, no covariate case (scenario A). Sample correlations together with posterior means (and 95% CSs in parenthesis) of estimated correlations.

cases as well. Note also the complete inadequacy of the latter in modeling positive dependence between SMF and LMF.

We then extended the model in two directions. In the first (scenario B), we added the logarithm of the size as covariate, to investigate whether big plants have a different biomass composition than small plants. In the second (scenario C), we investigated the dependence on time, and added one categorical covariate (namely the species) while neglecting the second (nitrate level). From Table 5, it emerges that the EFDReg and EFDReg$^{\mathrm{P}}$ models provide similar estimates of the parameters and similar WAIC values. Moreover, both models still recognize the presence of two (different) latent subpopulations (see Figure 6). In particular, the EFD-based group regressions detect the latent covariate (see Figure 6 where the model displays a quite remarkable fitting to high and low nitrate cases). As a consequence, the WAIC of the EFD models results in much smaller values than the DirReg one, and the estimate of the precision parameter is much higher. Similar remarks apply to scenario B. See Section 8.2 of the SM for a complete analysis of scenarios B and C.

Moreover, we considered the model used by Douma and Weedon, including the standardized time of harvest, its quadratic transformation, the logarithm of the size of the composition, and several double and triple interaction terms as covariates (scenario D). Furthermore, we included a regression on the precision parameter α^+, adopting a logarithm link function as illustrated in Section 3 and priors specified in Section 5 for regression coefficients $\boldsymbol{\gamma}$. We modeled the precision parameter adopting the same variables chosen by Douma and Weedon, namely species, nitrate supply, standardized time

		DirReg		EFDReg		EFDReg$^{\mathrm{P}}$	
	Parameter	Mean	95% CS	Mean	95% CS	Mean	95% CS
RMF	Intercept	**−0.038**	(−0.072, −0.004)	**−0.048**	(−0.081, −0.015)	**−0.070**	(−0.107, −0.033)
	Species	**0.426**	(0.375, 0.477)	**0.494**	(0.445, 0.544)	**0.508**	(0.460, 0.555)
	Time	**0.172**	(0.147, 0.197)	**0.146**	(0.122, 0.170)	**0.144**	(0.120 , 0.168)
SMF	Intercept	**−0.906**	(−0.950, −0.862)	**−0.916**	(−0.945, −0.888)	**−0.912**	(−0.942, −0.883)
	Species	**0.144**	(0.076, 0.212)	**0.135**	(0.089, 0.181)	**0.133**	(0.088, 0.178)
	Time	**0.064**	(0.031, 0.098)	**0.068**	(0.045, 0.091)	**0.068**	(0.046, 0.090)
	α^{+}	59.34	(54.19, 64.67)	124.91	(113.07, 137.25)	124.37	(112.67, 136.46)
	p_1	—	—	0.453	(0.426, 0.484)	0.386	(0.335, 0.439)
	p_2	—	—	0.547	(0.516, 0.574)	0.614	(0.561, 0.665)
	p_3	—	—	—	—	—	—
	w_1^N	—	—	0.236	(0.221, 0.252)	0.331	(0.309, 0.351)
	w_2^N	—	—	0.050	(0.024, 0.078)	0.058	(0.024, 0.091)
	w_3^N	—	—	—	—	—	—
	K	—		2 (100%)		2 (90.84%)	
	Mixture components	—		1 and 2 (99.995%)		1 and 2 (100%)	
	WAIC	−3061.8		−3372.2		−3365.7	

Table 5: Plants data with time and species as covariates (scenario C). Posterior means and 95% CSs of the parameters under the DirReg, EFDReg, and the EFDReg$^{\mathrm{P}}$ models. Posterior means with associated 95% CSs not containing the zero are rendered in bold. For the parameter K, the posterior mode, its posterior probability, and the chosen components are reported.

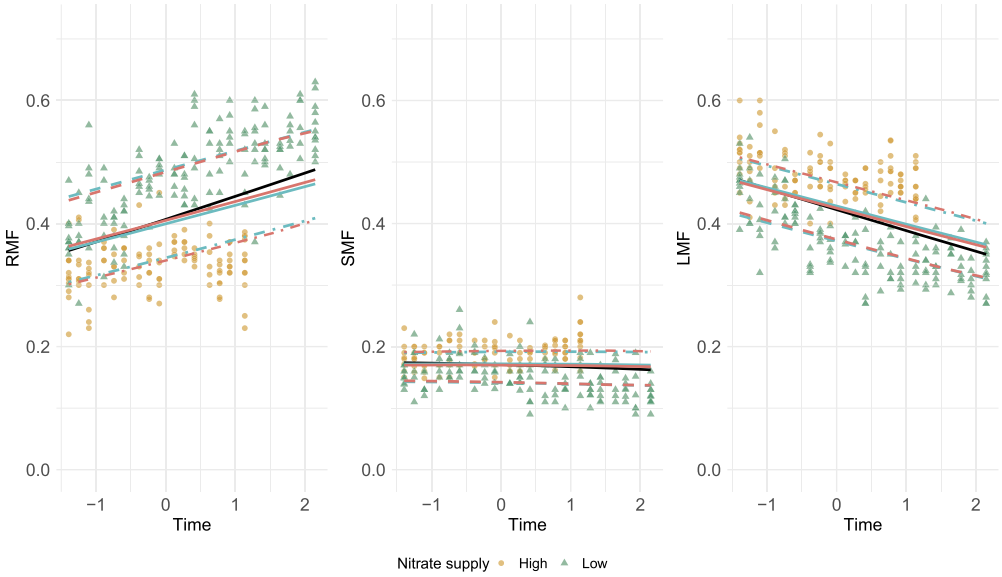

Figure 6: Plants data with time and species as covariates (scenario C): *D. flexuosa* plants. Fitted curves for the EFDReg (red), EFDReg$^{\mathrm{P}}$ (blue), and DirReg (black) models. The curves refer to parameters $\boldsymbol{\mu}$ (solid), $\boldsymbol{\lambda}_1$ (dashed), and $\boldsymbol{\lambda}_2$ (dot-dashed).

of harvest, and the interaction term between time and species. Table S14 shows that estimates of the regression coefficients are consistent across the three models. In particular, all the covariates introduced in the model provide a significant effect for at least one element of the composition or for the regression of the precision parameter. This is due to the variable selection performed by Douma and Weedon that we replicated. Both EFD-based regression models detect two latent groups of similar magnitude.

Figure 7 plots the predicted precision parameter α^+ as a function of the time of harvest, distinguishing on the basis of species and nitrate supply (we only report the EFDReg model as the EFDReg$^{\mathrm{P}}$ is nearly identical). Since the curves are very different and highly dependent on time, at least for *H. lanatus*, the importance of modeling α^+ as a function of covariates is evident. Finally, we note that, even under scenario D, where the presence of so many (properly selected) covariates makes likely the absence of residual latent groups, the estimated precision is uniformly higher in the EFDReg model than in the DirReg, leading to a better fitting.

We enriched the analysis of plants data by comparing the proposed models with respect to the models proposed by Graf (2020) and Ankam and Bouguila (2019). Since these competing models greatly differ from the DirReg and EFDReg models in terms of regression structure, and inferential approach, we based the comparison on the predictive ability of the models. Results are illustrated in Section 9 of the SM. Here we just remark that the EFDReg model shows the best predictive ability in all scenarios.

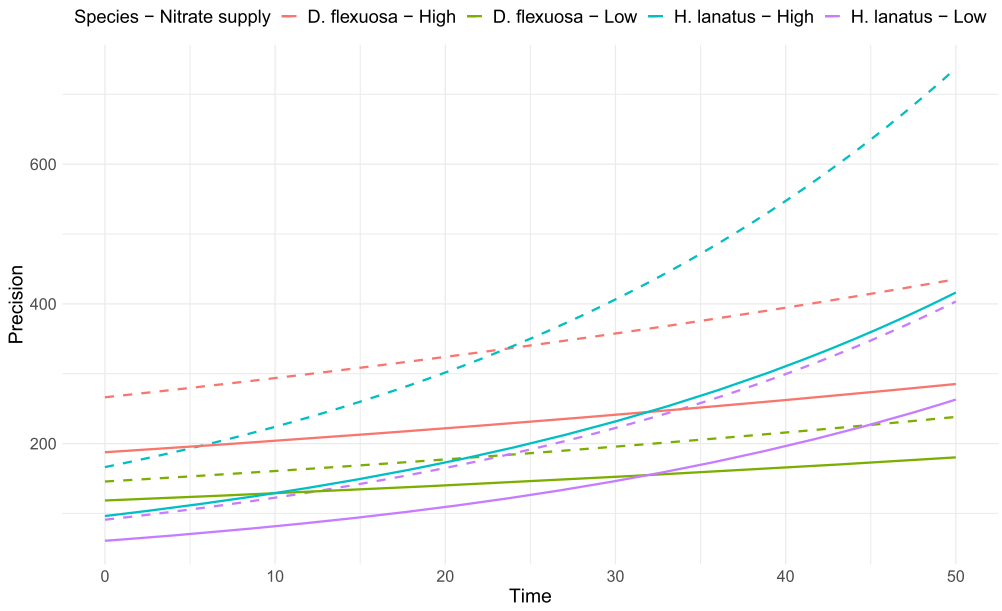

Figure 7: Plants data with all covariates (scenario D). Predicted precision under the DirReg (solid) and the EFDReg (dashed) models as function of time. Species and nitrate supply differ in color.

Finally, it is noteworthy that a numerical investigation (reported in Section 5 of the SM) showed robustness in the inferential conclusions with respect to different choices of the hyperparameter g appearing in the prior for the precision parameter α^+, the prior variance of the multivariate normal distributions for the regression coefficients, as well as with respect to alternative non- or weakly-informative priors.

8 Concluding remarks

Since both the EFDReg and EFDReg$^{\text{P}}$ models represent a clear improvement over the usual DirReg model, in future work we plan to extend them to further increase their flexibility, fitting ability, and interpretation of more complex data patterns. A relevant extension is the inclusion of random effects to allow for responses with a hierarchical structure (typically measured longitudinally or clustered), so that within-subject correlation can be handled. In some contexts the presence of values of the response on the support boundary (i.e., zero values for one or more elements) is certainly a relevant issue. Being continuous distributions, neither the Dirichlet nor the EFD can account for the presence of these values. Thus, it seems useful to study an "inflated" version of the EFDReg model obtained by introducing a discrete probabilistic scheme giving positive mass to zero values.

Note that, in principle, to achieve an even stronger flexibility one could consider a mixture of arbitrary Dirichlet regression models (i.e., where each component of the mixture has parameters completely unrelated to the other components' parameters). To the best of our knowledge this framework has not been explored in the literature. Though potentially interesting, this model faces severe estimation issues (going beyond the scope of this work) due to substantial identifiability problems generating label switching difficulties and unreliable posteriors, as we could experience even in simple simulation frameworks, as well as in application to real data. Furthermore, the general mixture type of regression model usually separately models the component regressions which display arbitrarily different dependencies on covariates. Therefore, unlike our model, it is unable to provide an easily interpretable evaluation of the general impact of covariates on the mean vector of the response variable.

As pointed to in Section 4, the potential of the EFDM distribution to analyze count data in a regression context seems worth exploring. The resulting model should be compared with the existing literature on the topic, which has recently become extensive especially with applications to microbiome data (e.g., see Chen and Li (2013); Subedi et al. (2020), and, for a Bayesian nonparametric approach, Ren et al. (2017)).

Finally, the use of the EFD as a prior for the weights of a generic mixture model could be a further issue to be tackled.

Supplementary Material

Supplementary Material of "A new multivariate regression model for bounded responses" (DOI: 10.1214/22-BA1359SUPP; .pdf). Section 1 describes the EFD tails' behavior. Sec-

tions 2 and 3 illustrate the derivation of the EFDReg parabolic and piecewise parabolic polynomial constraints, respectively. Section 4 provides proof of Proposition 1. Section 5 shows the results of the sensitivity analysis. Section 6 describes the general and simplified priors introduced for the cluster choice. Section 7 is devoted to a detailed description of the simulation studies. HMC Stan implementation, model diagnostic tools and comparison criteria can be found in Subsections 7.1 and 7.2. Subsections from 7.3 to 7.9 include a self-contained presentation of all simulation results. Section 8 includes a fully detailed description of the results concerning the application to plants data. Section 9 shows a brief comparison of our proposed model with competing models referring to plants data. Finally, Section 10 shows the performance of the proposed EFDReg model in the case of large values of D, through some simulation studies.

References

Aitchison, J. (2003). *The Statistical Analysis of Compositional data*. London: The Blackburn Press. MR0865647. doi: https://doi.org/10.1007/978-94-009-4109-0. 378, 392

Albert, J. (2009). *Bayesian computation with R*. London: Springer Science. MR2839312. doi: https://doi.org/10.1007/978-0-387-92298-0. 388

Ankam, D. and Bouguila, N. (2019). "Generalized Dirichlet regression and other compositional models with application to market-share data mining of information technology companies." In *Proceedings of the 21st International Conference on Enterprise Information Systems (ICEIS 2019)*, 158–166. 378, 383, 401

Ascari, R., Di Brisco, A. M., Migliorati, S., and Ongaro, A. (2023). "Supplementary Material for "A Multivariate Mixture Regression Model for Constrained Responses"." *Bayesian Analysis*. doi: https://doi.org/10.1214/22-BA1359SUPP. 379

Camargo, A., Stern, J., and Lauretto, M. (2012). "Estimation and Model Selection in Dirichlet Regression." In *Bayesian Inference and Maximum Entropy Methods in Science and Engineering, American Institute of Physics Conference Proceedings*, 206–213. 378, 383

Campbell, G. and Mosimann, J. E. (1987). "Multivariate analysis of size and shape: modelling with the Dirichlet distribution." In *ASA Proceedings of Section on Statistical Graphics*, 93–101. 378

Chen, J. and Li, H. (2013). "Variable selection for sparse Dirichlet-multinomial regression with an application to microbiome data analysis." *Annals of Applied Statistics*, 111(1): 418–442. URL https://doi.org/10.1214/12-AOAS592 MR3086425. doi: https://doi.org/10.1214/12-AOAS592. 402

Da-Silva, C. and Rodrigues, G. (2015). "Bayesian Dynamic Dirichlet Models." *Communications in Statistics—Simulation and Computation*, 44: 787–818. MR3257741. doi: https://doi.org/10.1080/03610918.2013.795592. 378, 383

Douma, J., and Weedon, J. (2019). "Analysing continuous proportions in ecology and

evolution: A practical introduction to beta and Dirichlet regression." *Methods in Ecology and Evolution*, 10: 1412–1430. 397

Frühwirth-Schnatter, S. (2006). *Finite mixture and Markov switching models*. Springer Science & Business Media. MR2265601. 384, 388

Gelman, A., Carlin, J., Stern, H., Dunson, D., Vehtari, A., and Rubin, D. (2013). *Bayesian Data Analysis*. London: CRCPress, third edition. MR3235677. 396

Graf, M. (2020). "Regression for compositions based on a generalization of the Dirichlet distribution." *Statistical Methods and Applications*, 29(4): 913–936. URL https://doi.org/10.1007/s10260-020-00512-y MR4174692. doi: https://doi.org/10.1007/s10260-020-00512-y. 383, 401

Gueorguieva, R., Rosenheck, R., and Zelterman, D. (2008). "Dirichlet component regression and its applications to psychiatric data." *Computational Statistics and Data Analysis*, 12(1): 5344–5355. MR2526600. doi: https://doi.org/10.1016/j.csda.2008.05.030. 378, 383

Hijazi, R. H. and Jernigan, R. W. (2009). "Modelling compositional data using Dirichlet regression models." *Journal of Applied Probability & Statistics*, 4(1): 77–91. MR2668780. 378, 383

Maier, M. (2014). "Dirichletreg: Dirichlet regression for compositional data in R." Technical Report Research Report Series, Department of Statistics and Mathematics, University of Economics and Business. 378, 383

McCullagh, P. and Nelder, J. (1989). *Generalized linear models*. London: Chapman & Hall. MR3223057. doi: https://doi.org/10.1007/978-1-4899-3242-6. 383

Migliorati, S., Ongaro, A., and Monti, G. (2017). "A structured Dirichlet mixture model for compositional data inferential and applicative issues." *Statistics and Computing*, 27: 963–983. MR3627557. doi: https://doi.org/10.1007/s11222-016-9665-y. 381

Morais, J., Thomas-Agnan, C., and Simioni, M. (2018). "Using compositional and Dirichlet models for market share regression." *Journal of Applied Statistics*, 45(9): 1670–1689. MR3800339. doi: https://doi.org/10.1080/02664763.2017.1389864. 378, 383

Neal, R. (1994). "An Improved Acceptance Procedure for the Hybrid Monte Carlo Algorithm." *Journal of Computational Physics*, 111(1): 194–203. MR1271540. doi: https://doi.org/10.1006/jcph.1994.1054. 388

Ongaro, A. and Migliorati, S. (2013). "A generalization of the Dirichlet distribution." *Journal of Multivariate Analysis*, 114(1): 412–426. MR2993896. doi: https://doi.org/10.1016/j.jmva.2012.07.007. 381

Ongaro, A., Migliorati, S., and Ascari, R. (2020). "A new mixture model on the simplex." *Statistics and Computing*, 30: 749–770. MR4108675. doi: https://doi.org/10.1007/s11222-019-09920-x. 378, 380, 381

Poorter, H., van de Vijver, C., Boot, R., and Lambers, H. (1995). "Growth and carbon economy of a fast-growing and a slow-growing grass species as dependent on nitrate supply." *Plant and Soil*, 171: 217–227. 397

Ren, B., Bacallado, S., Favaro, S., Vatanen, T., Huttenhower, C., and Trippa, L. (2017). "Bayesian Mixed Effects Models for Zero-inflated Compositions in Microbiome Data Analysis." *Annals of Applied Statistics*, 14(1): 494–517. URL http://dx.doi.org/10.1214/19-AOAS1295 MR4085103. doi: https://doi.org/10.1214/19-AOAS1295. 402

Stan Development Team (2016). *Stan Modeling Language Users Guide and Reference Manual*. URL http://mc-stan.org/ 391

Subedi, S., Neish, D., Bak, S., and Feng, Z. (2020). "Cluster analysis of microbiome data by using mixtures of Dirichlet-multinomial regression models." *Journal of the Royal Statistical Society. Series C (Applied Statistics)*, 69(5): 1163–1187. URL https://doi.org/10.1111/rssc.12432 MR4166861. doi: https://doi.org/10.1111/rssc.12432. 402

Tsagris, M. and Stewart, C. (2018). "A Dirichlet Regression Model for Compositional Data with Zeros." *Lobachevskii Journal of Mathematics*, 39(3): 398–412. MR3789428. doi: https://doi.org/10.1134/s1995080218030198. 378, 383

van der Merwe, S. (2019). "A method for Bayesian regression modelling of compositional data." *South African Statistical Journal*, 53(1): 55–64. MR3966355. 378, 383

Vehtari, A., Gelman, A., and Gabry, J. (2017). "Practical Bayesian model evaluation using leave-one-out cross-validation and WAIC." *Statistics and Computing*, 27(5): 1413–1432. MR3647105. doi: https://doi.org/10.1007/s11222-016-9696-4. 392

Watanabe, S. (2010). "Asymptotic equivalence of Bayes cross validation and widely applicable information criterion in singular learning theory." *Journal of Machine Learning Research*, 11: 3571–3594. MR2756194. 392

Acknowledgments

We greatly acknowledge the DEMS Data Science Lab for supporting this work by providing computational resources. The second author is member of the "Gruppo Nazionale per l'Analisi Matematica, la Probabilità e le loro Applicazioni" (GNAMPA) of the "Istituto Nazionale di Alta Matematica" (INdAM).