Submitted to the Annals of Applied Statistics

A LATENT VARIABLE MIXTURE MODEL FOR COMPOSITION-ON-COMPOSITION REGRESSION WITH APPLICATION TO CHEMICAL RECYCLING

BY NICHOLAS RIOS[1,a], LINGZHOU XUE[2,b] AND XIANG ZHAN[3,c]

[1]*Department of Statistics, George Mason University,* [a]*nrios4@gmu.edu*

[2]*Department of Statistics, Pennsylvania State University,* [b]*lzxue@psu.edu*

[3]*Department of Biostatistics and Beijing International Center for Mathematical Research, Peking University,* [c]*zhanx@bjmu.edu.cn*

It is quite common to encounter compositional data in a regression framework in data analysis. When both responses and predictors are compositional, most existing models rely on a family of log-ratio based transformations to move the analysis from the simplex to the reals. This often makes the interpretation of the model more complex. A transformation-free regression model was recently developed, but it only allows for a single compositional predictor. However, many datasets include multiple compositional predictors of interest. Motivated by an application to Hydrothermal Liquefaction (HTL) data, a novel extension of this transformation-free regression model is provided that allows for two (or more) compositional predictors to be used via a latent variable mixture. A modified Expectation-Maximization algorithm is proposed to estimate model parameters, which are shown to have natural interpretations. Conformal inference is used to obtain prediction limits on the compositional response. The resulting methodology is applied to the HTL dataset. Extensions to multiple predictors are discussed.

1. Introduction. As a motivating example, we consider a recent dataset on hydrothermal liquefaction (HTL) in chemical recycling. At high temperatures and pressures, biomass (e.g. plastic) undergoes reactions that produce bio-oil. This allows for materials that would otherwise be discarded as waste to have some use as a future energy source. The response is the composition of the produced oil, which has nine dimensions. The dataset also provides the chemical and elemental compositions of the original biomass, which have four dimensions each. The uses of the resulting oil from the HTL process are determined by the composition of the oil, as the oil composition influences acidity, octane number, heating value, and other important chemical properties (Gollakota, Kishore and Gu, 2018). Existing approaches for predicting the oil composition are time-consuming and involve solving a customized set of simultaneous differential equations, which would have many parameters for the diverse HTL process (Valdez, Tocco and Savage, 2014). In existing literature, statistical models have been developed that use biochemical or elemental compositions to predict oil yield, but not oil composition (Lu et al., 2018). The HTL process is a highly complex and interconnected thermochemical process with underlying chemical reactions that are not fully understood (Shahbeik et al., 2024). Guirguis et al. (2024) noted that one source of this complexity is the diversity that exists in the components of the biomass, which makes modeling direct reaction pathways for each individual biomass incredibly challenging. There is evidence that certain properties of the biomass make one type of composition more informative than the other. For example, McKendry (2002) notes that the ratio of cellulose and lignin are critical for wet biomass conversion, while elemental components (primarily carbon) are more critical

Keywords and phrases: Compositional data, EM-algorithm, Kullback-Leibler distance, Transformation-free.

for dry biomass. When studying the oil yield of lignocellulose biomass, Zhang et al. (2021) noted that the elemental composition of the initial biomass affected the yield, and claimed the carbon chain number was a latent factor that influenced this result. Since the underlying nature of the HTL process is highly complex, we make a simplifying assumption that the oil composition is a latent mixture of chemical and elemental compositions. This assumption not only helps us develop an interpretable model, but it also helps us account for heterogeneity of the biomasses in the examined HTL dataset. Table 1 shows the distribution of biomasses by category in the HTL dataset. As the table shows, the dataset covers many types of biomasses, including macroalgae, microalgae, manure, sawdust, sewage sludge, and scrap tires.

Algae (Macro)	Straw, Grain	Algae (Micro)	Other	Grass, Shrub	Food Waste	Wood, Sawdust	Manure	Sewage Sludge	Scrap Tire
0.27	0.18	0.16	0.16	0.06	0.04	0.07	0.03	0.02	0.01

TABLE 1
Distribution of Biomasses in HTL Dataset

Regression analysis has been frequently used for inference of problems involving compositional data, and specifically, regression models with compositional predictors has mainly been done via log ratio transformations. This concept was first explored by Aitchison and Bacon-Shone (1984), who proposed a linear log-contrast procedure for hypothesis tests on compositional covariates. Lin et al. (2014) adopted a linear log-contrast model in a high-dimensional setting and minimized an ℓ_1 regularization criterion to estimate the parameters. Srinivasan, Xue and Zhan (2022) considered the analysis of high-dimensional microbiome compositional data and studied knockoff filters to identify significant microbiome features in a scalar-on-composition regression framework. In this context, the compositional data were predictors for a scalar response of interest. On the other hand, when handling compositional responses, it is possible to first swap the role of responses and predictors and then use these aforementioned log ratio-based transformations in the corresponding inverse regression model. Moreover, some commonly studied strategies were Dirichlet regression (Hijazi and Jernigan, 2009) and Dirichlet-multinomial regression (Mosimann, 1962), which first modeled the distribution of compositional responses via a Dirichlet-based distribution, and then linked key parameters of the underlying distribution to non-compositional predictors or covariates of interest (Chen and Li, 2013; Douma and Weedon, 2019; Tang and Chen, 2019). Finally, composition-on-composition regression has been studied through a transformation-based model framework. Chen, Zhang and Li (2017) applied an isometric log ratio (ilr) transformation to compositional predictors and a compositional response in a regression model.

Despite the large volume of research on compositional data, there are few transformation-free techniques for composition-on-composition regression. Transformation-based methods often make use of log ratio transformations; however, these transformations are of limited use when there are zeros in the predictor and the response. In the motivating example, the HTL process dataset includes many cases where some of the components in the composition are zero in both the predictors and the response. Furthermore, log-transformation models are often more difficult to interpret directly and some may require graphs to interpret main effects. See Section 2 for more details on existing transformation-based methods. All these have limited applications of log transformation-based methods in compositional data analysis. Recently, Fiksel, Zeger and Datta (2022) proposed a transformation-free composition-on-composition regression model for a single compositional predictor and response. This approach provided intuitive interpretations of the regression parameters, and did not require any strong distributional assumptions on the compositional response. However, it was limited in that it did not generalize to allow multiple compositional predictors, or allow for the inclusion of continuous predictors in the model. In particular, the model from Fiksel, Zeger and Datta (2022) did not address the motivating example because it treated its predictors as a single composition that must sum to 1, and did not allow for two compositions or subcompositions without a significant change to model interpretation.

In this paper, we propose a flexible composition-on-composition latent variable model that can handle mixtures of two (or more) compositional predictors. We first review existing methods of regression analysis for compositional predictors in Section 2. In Section 3, we first introduce a general framework for a latent variable regression model that models the response as a mixture (or convex combination) of two conditional distributions, with parameters selected based on the minimization of the Kullback Leibler distance. An EM algorithm is then derived for model estimation and the proposed methodology is extended to three or more predictor variables. Statistical inference procedures are also investigated in Section 3.4. Comprehensive simulation studies are conducted in Section 4 to evaluate the performance of the proposed methodology. In Section 5, the proposed methodology is applied to the HTL dataset, with results shown. Section 6 concludes the paper.

2. Preliminaries.

This section provides a brief overview of existing methods for regression analysis of compositional data. First, transformation-based methods are reviewed in Subsection 2.1, as the majority of existing research in this area uses some form of transformation. Then, recent transformation-free results are summarized in Subsection 2.2.

2.1. *Transformation-Based Methods.*

The sum-to-one nature of compositional data renders many classic statistical methods inappropriate or inadequate for compositional data analysis. For example, it is well known that spurious associations would be found in compositional data analysis if ignoring compositionality (Pearson, 1897). Many log-ratio transformations or log-contrast models for compositional regression have been explored to address this analysis challenge. Two of the first transformation methods explored were the additive log-ratio (ALR) and centered log-ratio (CLR) transformations (Aitchison, 1982). For a compositional predictor $\mathbf{z} \in \{(z_1, \ldots, z_D) \mid \sum_{i=1}^{D} z_i = 1, z_i \in (0,1)\}$, the ALR and CLR transformations are

$$\text{alr}(\mathbf{z})_j = \log\Big(\frac{z_j}{z_D}\Big), \tag{1}$$

where $j = 1, \ldots, D-1$, and

$$\text{clr}(\mathbf{z})_j = \log\Big(\frac{z_j}{(\prod_{i=1}^{D} z_j)^{1/D}}\Big), \tag{2}$$

where $j = 1, \ldots, D$. Another popular method is the isometric log-ratio (ILR) transformation (Egozcue et al., 2003), which is given by

$$\text{ilr}(\mathbf{z})_j = \sqrt{\frac{D-j}{D-j+1}} \log\Big(\frac{z_j}{(\prod_{k=j+1}^{D} z_k)^{1/(D-j)}}\Big) \tag{3}$$

for $j = 1, \ldots, D-1$.

These log-ratios are widely used in downstream statistical modeling of compositional data analysis and the major draw of these transformations is that, once the compositional data are transformed, they can be used in standard linear regression models without concern about the sum-to-one constraint. By adapting statistical methods on these transformed features, many versatile statistical methods for compositional data analysis have been proposed (Aitchison, 1982; Lin et al., 2014; Srinivasan, Xue and Zhan, 2021).

While these transformations-based methods first map the simplex $\mathcal{S}^D$ to the reals and then adapt classic versatile statistical methods to these transformed data, a key limitation is that, they do not accommodate compositional data with zeros and ones, which have been widely observed in the field. Moreover, they make results difficult to interpret. This often forces analysts to rely on graphs to display how the response changes with the compositional inputs, but graphs may not be easy to present when the dimension is large. This motivates the study of transformation-free methods for compositional regression.

2.2. *Transformation-Free Methods.* Fiksel, Zeger and Datta (2022) proposed a novel transformation-free direct regression model for a single compositional predictor $\mathbf{x} \in \mathcal{S}^{D_s}$ and response $\mathbf{y} \in \mathcal{S}^{D_r}$. The proposed method is a linear model of the form

$$E[\mathbf{y} \mid \mathbf{x}] = \mathbf{B}^T \mathbf{x} \tag{4}$$

where $\mathbf{B} \in \{\mathbb{R}^{D_s \times D_r} \mid B_{jk} \geq 0, \sum_{k=1}^{D_r} B_{jk} = 1 \text{ for } j = 1, \ldots, D_s\}$. This implies that $\mathbf{B}$ is a transition (Markov) matrix where each row sums to one and has nonnegative entries. These restrictions on $\mathbf{B}$ ensure that the predicted response will be compositional. The entries of $\mathbf{B}$ give a direct interpretation of how changing the composition of $\mathbf{x}$ affects the expected composition of $\mathbf{y}$. Let x_j and x_k be the j^{th} and k^{th} components of $\mathbf{x}$, and let $\mathbf{B}_{j*}, \mathbf{B}_{k*}$ denote the j^{th} and k^{th} rows of $\mathbf{B}$, respectively. If x_j increases by some small Δ and x_k $(k \neq j)$ decreases by the same Δ (since the components must sum to one), then the change in $E(\mathbf{y})$ is $\Delta(\mathbf{B}_{j*} - \mathbf{B}_{k*})$, assuming all other proportions in $\mathbf{x}$ are held constant. While the authors called their method direct regression (without transformations) (Fiksel, Zeger and Datta, 2022), we will refer to any method that regresses one compositional vector on at least another compositional vector (as done in Equation (4)) as composition-on-composition regression hereafter in this paper.

To estimate the parameters in $\mathbf{B}$, an Expectation-Maximization (EM) algorithm was used to minimize a loss function (ℓ) that is based on the Kullback-Leibler distance (KLD). Let x_{ij} be the j^{th} component of $\mathbf{x}_i$, y_{ik} be the k^{th} component of $\mathbf{y}_i$, and B_{jk} be the k^{th} entry of row j of $\mathbf{B}$. Specifically, the optimization problem is

$$\min_{\mathbf{B}} \ell(B; X, Y) = \min_{\mathbf{B}} - \sum_{i=1}^{N} \sum_{k=1}^{D_r} y_{ik} \log\Big(\frac{\sum_{j=1}^{D_s} B_{jk} x_{ij}}{y_{ik}}\Big) = \max_{\mathbf{B}} \sum_{i=1}^{N} \sum_{k=1}^{D_r} y_{ik} \log\Big(\sum_{j=1}^{D_s} B_{jk} x_{ij}\Big). \tag{5}$$

The maximization at the final step of (5) is performed using the EM algorithm. The implementation of this algorithm is available in the R package codalm. Since closed forms for the $E-$ and $M-$ steps of this algorithm are provided, the implementation is very fast.

3. Methodology.

In this section, the methodology for the latent variable composition-on-composition regression model is introduced, and then a modified EM algorithm is proposed to estimate its parameters. This model is then extended to accommodate three or more compositional predictors, and procedures for model inference are investigated in the end.

3.1. *A Latent Variable Mixture Model.* Assume that the data follow a latent variable mixture model

$$f(\mathbf{y}_i \mid w_i^*, \mathbf{x}_{1i}, \mathbf{x}_{2i}) = w_i^* f_1(\mathbf{y}_i \mid \mathbf{x}_{1i}) + (1 - w_i^*) f_2(\mathbf{y}_i \mid \mathbf{x}_{2i}) \tag{6}$$

$$w_i^* \sim \text{Bernoulli}(\theta) \tag{7}$$

for $i = 1, \ldots, N$, where $\mathbf{x}_{1i}, \mathbf{x}_{2i}$ are compositional vectors for the i^{th} observation (of dimensions D_{s_1}, D_{s_2}, respectively), and $\mathbf{y}_i$ is the compositional response (of dimension D_r) for the i^{th} observation. Also, f_1, f_2 are component densities of $\mathbf{y}_i \mid \mathbf{x}_{1i}, \mathbf{y}_i \mid \mathbf{x}_{2i}$, respectively. The latent variables w_i^* determine if the response $\mathbf{y}_i$ depends on the first or second compositional predictor. Then, from the law of iterated expectations, it directly follows that

$$E[\mathbf{y}_i \mid \mathbf{x}_{1i}, \mathbf{x}_{2i}] = E[E[\mathbf{y}_i \mid w_i^*, \mathbf{x}_{1i}, \mathbf{x}_{2i}]] = \theta E_{f_1}[\mathbf{y}_i \mid \mathbf{x}_{1i}] + (1-\theta) E_{f_2}[\mathbf{y}_i \mid \mathbf{x}_{2i}] \tag{8}$$

The proposed model focuses only on the first moment $E[\mathbf{y}_i \mid \mathbf{x}_{1i}, \mathbf{x}_{2i}]$, which is parameterized as follows:

$$E[\mathbf{y}_i \mid \mathbf{x}_{1i}, \mathbf{x}_{2i}] = \theta \mathbf{B}_1^T \mathbf{x}_{1i} + (1-\theta) \mathbf{B}_2^T \mathbf{x}_{2i} \tag{9}$$

where $\theta \in (0,1)$, and $\mathbf{B}_1, \mathbf{B}_2$ are Markov (transition) matrices of dimensions $D_{s_1} \times D_r, D_{s_2} \times D_r$, respectively, and D_r is the dimension of the response $\mathbf{y}_i$. Since $\mathbf{B}_1, \mathbf{B}_2$ are Markov matrices, then their entries must be nonnegative values such that each of their rows sum to 1. Since $\mathbf{B}_1^T \mathbf{x}_1$ and $\mathbf{B}_2^T \mathbf{x}_2$ lie in the response simplex, then any convex combination of them must also lie in the same simplex. To estimate the parameters, we consider minimizing the Kullback-Leibler distance (KLD) between each $\mathbf{y}_i$ and $E[\mathbf{y}_i \mid \mathbf{x}_{1i}, \mathbf{x}_{2i}]$.

$$\hat{B}_1, \hat{B}_2, \hat{\theta} = \underset{\mathbf{B}_1, \mathbf{B}_2, \theta}{\arg\min} - \sum_{i=1}^{N} \sum_{\ell=1}^{D_r} y_{i\ell} \log \Big(\frac{\theta \sum_{j=1}^{D_{s_1}} B_{1j\ell} x_{1ij} + (1-\theta) \sum_{k=1}^{D_{s_2}} B_{2k\ell} x_{2ik}}{y_{i\ell}} \Big) \tag{10}$$

$$= \underset{\mathbf{B}_1, \mathbf{B}_2, \theta}{\arg\max} \sum_{i=1}^{N} \sum_{\ell=1}^{D_r} y_{i\ell} \log \Big(\theta \sum_{j=1}^{D_{s_1}} B_{1j\ell} x_{1ij} + (1-\theta) \sum_{k=1}^{D_{s_2}} B_{2k\ell} x_{2ik} \Big) \tag{11}$$

$$= \underset{\mathbf{B}_1, \mathbf{B}_2, \theta}{\arg\max} \sum_{i=1}^{N} \sum_{\ell=1}^{D_r} y_{i\ell} \log \Big(\sum_{j=1}^{D_{s_1}} \sum_{k=1}^{D_{s_2}} \big[\theta B_{1j\ell} x_{1ij} x_{2ik} + (1-\theta) B_{2k\ell} x_{1ij} x_{2ik} \big] \Big), \tag{12}$$

where $y_{i\ell}$ denotes the ℓ^{th} component of $\mathbf{y}_i$, x_{1ij} denote the j^{th} component of $\mathbf{x}_{1i}$, x_{2ik} denotes the k^{th} component of $\mathbf{x}_{2i}$, $B_{1j\ell}$ is the element of $\mathbf{B}_1$ at row j and column ℓ, and $B_{2k\ell}$ is similarly defined. The equality in (12) exploits the fact that the predictors $\mathbf{x}_1, \mathbf{x}_2$ are compositional, i.e. $\sum_{j=1}^{D_{s_1}} x_{1ij} = 1$ and $\sum_{k=1}^{D_{s_2}} x_{2ik} = 1$. The parameters $\theta, \mathbf{B}_1, \mathbf{B}_2$ will be estimated using an EM algorithm described in the subsequent sections.

It should be noted that the optimization in (12) can be viewed from an estimating equation approach. Denote the vector of model parameters as $\mathbf{\Omega} = (\mathbf{B}_1, \mathbf{B}_2, \theta)$. Then, we want a function ℓ that satisfies

$$E\Big[\frac{d\ell}{d\mathbf{\Omega}} \mid \mathbf{\Omega}_0\Big] = 0 \tag{13}$$

where $\mathbf{\Omega}_0$ are the true parameters. The function ℓ is taken to be the KLD, which is similar to Fiksel, Zeger and Datta (2022). The approach taken in this paper will reduce to that of Fiksel, Zeger and Datta (2022) if the parameter space is restricted to the case where $\theta = 0$ or 1 (i.e., only one compositional predictor).

3.2. *A Modified EM Algorithm.*

The EM algorithm is often used to estimate unknown parameters in mixture models (McLachlan and Krishnan, 2007) with recent applications to sports marketing (DeSarbo, Chen and Blank, 2017), network psychometrics (Lee and Xue, 2018; Lee et al., 2022), water pollution analysis (Agarwal and Xue, 2020), and others. In this section, a modified EM algorithm is provided to estimate the parameters $\theta, \mathbf{B}_1, \mathbf{B}_2$ in Model (9). Fiksel, Zeger and Datta (2022) derived an EM algorithm in the single-predictor case under a multinomial assumption. We have provided similar derivations that follow their work for the case of two predictors in Section S1 of the supplementary materials (Rios, Xue and Zhan, 2024) and we now extend it to the general case that responses are compositional. It has been shown that, at iteration t of the EM algorithm, the $E-$step requires computing the quantities

$$\pi_{1ij\ell}^{(t+1)} = \frac{\theta^{(t)} x_{1ij} B_{1j\ell}^{(t)}}{\theta^{(t)} \sum_{j=1}^{D_{s_1}} x_{1ij} B_{1j\ell}^{(t)} + (1-\theta^{(t)}) \sum_{k=1}^{D_{s_2}} x_{2ik} B_{2k\ell}^{(t)}}, \tag{14}$$

$$\pi_{2ik\ell}^{(t+1)} = \frac{(1-\theta^{(t)}) x_{2ik} B_{2k\ell}^{(t)}}{\theta^{(t)} \sum_{j=1}^{D_{s_1}} x_{1ij} B_{1j\ell}^{(t)} + (1-\theta^{(t)}) \sum_{k=1}^{D_{s_2}} x_{2ik} B_{2k\ell}^{(t)}}. \tag{15}$$

The $M-$step for updating the entries of $\mathbf{B}_1, \mathbf{B}_2$ is then

$$B_{1j\ell}^{(t+1)} = \frac{\sum_{i=1}^{N} y_{i\ell} \pi_{1ij\ell}^{(t+1)}}{\sum_{i=1}^{N} \sum_{\ell=1}^{D_r} y_{i\ell} \pi_{1ij\ell}^{(t+1)}}, \tag{16}$$

$$B_{2k\ell}^{(t+1)} = \frac{\sum_{i=1}^{N} y_{i\ell} \pi_{2ik\ell}^{(t+1)}}{\sum_{i=1}^{N} \sum_{\ell=1}^{D_r} y_{i\ell} \pi_{2ik\ell}^{(t+1)}}. \tag{17}$$

While it is possible to also derive similar updates for θ under the multinomial assumption, it is not trivial to do so in general when the distribution of the data is unknown. However, (16) and (17) imply that, given a fixed value of $\theta^{(t)}$, one can update the entries of $\mathbf{B}_1$ and $\mathbf{B}_2$. Inspired by this idea, a modified EM algorithm is proposed to estimate $\mathbf{B}_1, \mathbf{B}_2$, and θ. For ease of notation, let $\mathbf{X}_1 = \{\mathbf{x}_{1i}, i = 1, \ldots, N\}$, $\mathbf{X}_2 = \{\mathbf{x}_{2i}, i = 1, \ldots, N\}$, $\mathbf{Y} = \{\mathbf{y}_i, i = 1, \ldots, N\}$, and $\text{KLD}(\mathbf{Y} || E[\mathbf{Y}|\mathbf{X}_1, \mathbf{X}_2]) = \sum_{i=1}^{N} \text{KLD}(\mathbf{y}_i || E[\mathbf{y}_i | \mathbf{x}_{1i}, \mathbf{x}_{2i}])$.

Algorithm 1: Modified EM Algorithm for Dual Composition Model

```
Inputs: Data $\mathbf{X}_1, \mathbf{X}_2, \mathbf{Y}$, initial $\mathbf{B}_1^{(0)}, \mathbf{B}_2^{(0)}, \theta^{(0)}$
1. Set $\delta = \infty, t = 0$, and $\text{KLD}_0 = \text{KLD}(\mathbf{Y} \,||\, E[\mathbf{Y} \mid \mathbf{X}_1, \mathbf{X}_2, \mathbf{B}_1^{(0)}, \mathbf{B}_2^{(0)}, \theta^{(0)}])$.
2. while $\delta > 10^{-8}$ do
    3. E-step. Compute $\pi_{1ij\ell}, \pi_{2ik\ell}$ in Equations (14), (15) using the entries of
       $\mathbf{B}_1^{(t)}, \mathbf{B}_2^{(t)}$ for each $i = 1, \ldots, N$, $j = 1, \ldots, D_{s_1}$, $k = 1, \ldots, D_{s_2}$, $\ell = 1, \ldots, D_r$
       using the current value of $\theta^{(t)}$.
    4. M-step. Use equations (16) and (17) to find the entries of $\mathbf{B}_1^{(t+1)}, \mathbf{B}_2^{(t+1)}$.
    5. for each $\theta$ over a fine grid on $(0, 1)$ do
        6. Compute and store $\text{KLD}_\theta = \text{KLD}(\mathbf{Y} \,||\, E[\mathbf{Y} \mid \mathbf{X}_1, \mathbf{X}_2, \mathbf{B}_1^{(t+1)}, \mathbf{B}_2^{(t+1)}, \theta])$.
    end
    7. $\theta^{(t+1)} = \arg\min_\theta \text{KLD}_\theta$. Store the smallest KLD as $\text{KLD}_{t+1}$.
    8. Update $\delta = |\text{KLD}_{t+1} - \text{KLD}_t|$, and update $t = t + 1$.
end
Output $\hat{\mathbf{B}}_1 = \mathbf{B}_1^{(t)}, \hat{\mathbf{B}}_2 = \mathbf{B}_2^{(t)}, \hat{\theta} = \theta^{(t)}$
```

As inputs, Algorithm 1 takes compositional predictors of dimension $N \times D_{s_1}, N \times D_{s_2}$, compositional responses of dimension $N \times D_r$, and initial values for $\mathbf{B}_1, \mathbf{B}_2$, and θ. At every t^{th} iteration, the EM algorithm is used to find the values $\mathbf{B}_1^{(t+1)}$ and $\mathbf{B}_2^{(t+1)}$ that minimize the KLD for the previous $\theta^{(t)}$. Then, $\theta^{(t+1)}$ is set to be the value that minimizes the Kullback-Leibler Distance between the actual response and the expected response (given $\mathbf{B}_1^{(t+1)}$ and $\mathbf{B}_2^{(t+1)}$) over a grid of θ values in $(0, 1)$. Since the EM updates for $\mathbf{B}_1$ and $\mathbf{B}_2$ have a closed form, and it is easy to compute the KLD, these updates can be performed quickly. The algorithm converges when the absolute difference between successive KLDs falls below a small tolerance. Corollary 1 states that the updated estimates in Steps 3 and 4 of each iteration of Algorithm 1 will increase the target multinomial quasi-likelihood for fixed θ, and, equivalently, decrease the KLD when **y** is compositional. Therefore, the EM algorithm will converge. The proof of Corollary 1 can be found in Section S2 of the supplementary materials (Rios, Xue and Zhan, 2024).

COROLLARY 1. *Let* $f(t) = \sum_{i=1}^{N}\sum_{\ell=1}^{D_r} y_{i\ell}\log\Big(\sum_{j=1}^{D_{s_1}}\sum_{k=1}^{D_{s_2}}\big[\theta B_{1j\ell}^{(t)}x_{1ij}x_{2ik} + (1-\theta)B_{2k\ell}^{(t)}x_{1ij}x_{2ik}\big]\Big)$ *be the value of the objective function at iteration* t *of the EM algorithm described in Steps 3 and 4 of Algorithm 1 for a fixed* θ *when the response* **y** *is compositional. Then* $f(t+1) - f(t) \geq 0$.

3.3. *Three or More Compositional Predictors.* The latent variable model with two predictors (9) has a natural extension to three or more compositional predictors. Suppose there are p predictor compositions of interest $\mathbf{x}_1, \ldots, \mathbf{x}_p$. Then the model of interest is

$$E[y \mid \mathbf{x}_1, \ldots, x_p] = \theta_1 \mathbf{B}_1^T \mathbf{x}_1 + \theta_2 \mathbf{B}_2^T \mathbf{x}_2 + \cdots + \theta_p \mathbf{B}_p^T \mathbf{x}_p \tag{18}$$

where $\theta_1 + \theta_2 + \cdots + \theta_p = 1$, and each $\mathbf{B}_s$ is a transition (Markov) matrix for $s = 1, \ldots, p$. In other words, the latent variable model is a convex combination of the points $\mathbf{B}_s^T\mathbf{x}_s \in \mathcal{S}^{D_r}$ for $s = 1, \ldots, p$. This ensures that the predicted response is a valid point in the response simplex. In this model, one may re-parameterize $\theta_p = 1 - \sum_{s=1}^{p-1}\theta_s$.

To estimate the parameters for Model (18), Algorithm 1 can be run over a multi-dimensional grid for $\boldsymbol{\theta} = (\theta_1, \ldots, \theta_{p-1})$. The E-step becomes

$$\pi_{sij\ell}^{(t+1)} = \frac{\theta_s^{(t)} x_{sij} B_{sj\ell}^{(t)}}{\sum_{s=1}^{p-1}[\theta_s^{(t)} \sum_{j=1}^{D_s} x_{sij} B_{sj\ell}^{(t)}] + (1 - \sum_{s=1}^{p-1}\theta_s^{(t)})\sum_{k=1}^{D_s} x_{pik} B_{pk\ell}^{(t)}} \tag{19}$$

and the M-step becomes

$$B_{sj\ell}^{(t+1)} = \frac{\sum_{i=1}^{N} y_{i\ell}\pi_{sij\ell}^{(t+1)}}{\sum_{i=1}^{N}\sum_{\ell=1}^{D_r} y_{i\ell}\pi_{sij\ell}^{(t+1)}} \tag{20}$$

where $j = 1, \ldots, \dim(x_s)$. The theoretical results of Corollary 1 can naturally be extended to higher dimensions. This is summarized in Corollary 2.

COROLLARY 2. *Let* $f(t) = \sum_{i=1}^{N}\sum_{\ell=1}^{D_r} y_{i\ell}\log\Big(\sum_{s=1}^{p}\sum_{j=1}^{D_s}\big[\theta_s B_{sj\ell}^{(t)} x_{sij}\big]\Big)$ *be the value of the objective function at iteration* t *of the EM algorithm described in equations (19) and (20) for fixed* $\theta_1, \ldots, \theta_p$*, such that* $\sum_{s=1}^{p}\theta_p = 1$ *and* $\theta_s > 0$*, when the response* **y** *is compositional. Then* $f(t+1) - f(t) \geq 0$.

Using the results of Corollary 2, the EM algorithm can be modified for $p \geq 3$ compositional predictors. The proof of Corollary 2 can be found in Section S2 of the supplementary materials (Rios, Xue and Zhan, 2024). The modified algorithm is given in Algorithm 2.

Algorithm 2: Modified EM Algorithm for Multiple Composition Model

```
Inputs: Data $\mathbf{X}_1, \mathbf{X}_2, \ldots, \mathbf{X}_p, \mathbf{Y}$, initial $\mathbf{B}_1^{(0)}, \mathbf{B}_2^{(0)}, \ldots, \mathbf{B}_p^{(0)} \boldsymbol{\theta}^{(0)}$
1. Set $\delta = \infty, t = 0$, and $\mathrm{KLD}_0 = \mathrm{KLD}(\mathbf{Y} \,||\, E[\mathbf{Y} \mid \mathbf{X}_1, \mathbf{X}_2, \mathbf{B}_1^{(0)}, \mathbf{B}_2^{(0)}, \boldsymbol{\theta}^{(0)}])$.
2. while $\delta > 10^{-8}$ do
    3. E-step. Compute the conditional expectations $\pi_{1ij\ell}, \ldots \pi_{pij\ell}$ in Equation (19)
       using the entries of $\mathbf{B}_1^{(t)}, \mathbf{B}_2^{(t)}, \ldots, \mathbf{B}_p^{(t)}$ for each $i = 1, \ldots, N$, $j = 1, \ldots, D_s$,
       $\ell = 1, \ldots, D_r$, using the current vector $\boldsymbol{\theta}^{(t)}$.
    4. M-step. Use equation (20) to find the entries of $\mathbf{B}_1^{(t+1)}, \mathbf{B}_2^{(t+1)}, \ldots, \mathbf{B}_p^{(t+1)}$.
    5. for each $\boldsymbol{\theta}$ over a fine grid on $\mathcal{S}^{p-1}$ do
        6. Compute and store $\mathrm{KLD}_{\boldsymbol{\theta}} = \mathrm{KLD}(\mathbf{Y} \,||\, E[\mathbf{Y} \mid \mathbf{X}_1, \mathbf{X}_2, \mathbf{B}_1^{(t+1)}, \mathbf{B}_2^{(t+1)}, \boldsymbol{\theta}])$.
    end
    7. $\boldsymbol{\theta}^{(t+1)} = \arg\min_{\boldsymbol{\theta}} \mathrm{KLD}_{\boldsymbol{\theta}}$. Store the smallest KLD as $\mathrm{KLD}_{t+1}$.
    8. Update $\delta = |\mathrm{KLD}_{t+1} - \mathrm{KLD}_t|$, and update $t = t + 1$.
end
Output $\hat{\mathbf{B}}_1 = \mathbf{B}_1^{(t)}, \hat{\mathbf{B}}_2 = \mathbf{B}_2^{(t)}, \hat{\boldsymbol{\theta}} = \boldsymbol{\theta}^{(t)}$
```

To implement Algorithm 2, the grid would need to be over a $(p-1)$ dimensional simplex. For $p = 2$, this reduces to a fine grid over the unit interval $(0, 1)$, which is precisely what is done in Algorithm 1. For $p = 3$, a fine grid over the unit simplex would be needed, which is computationally feasible. For large p, the curse of dimensionality makes it costly to perform a grid search to find the optimal $\boldsymbol{\theta}$ and grid search may not be the best approach in this case. Finding optimal parameters in this high-dimensional setting is an open research area.

3.4. *Model Inference.* After finding estimates for the model parameters $\mathbf{B}_1, \mathbf{B}_2$, and θ, the next logical step is to quantify the uncertainty about these values and perform statistical inference. Since Model (9) does not specify an exact distribution for the compositional response, estimating the uncertainty associated with the estimation of model parameters can be done in a nonparametric manner. In particular, 95% confidence intervals for $\theta, \mathbf{B}_1$ and $\mathbf{B}_2$ can be constructed using bootstrapping. To obtain a bootstrap sample, the rows of original data $(\mathbf{x}_{1i}, \mathbf{x}_{2i}, \mathbf{y}_i)$ are resampled with replacement. Then, estimates of $\mathbf{B}_1, \mathbf{B}_2$, and θ can be calculated using this bootstrap sample. This procedure is repeated a relatively large number of times. Then, for each element of $\mathbf{B}_1, \mathbf{B}_2$, and for θ, 95% confidence intervals can be obtained using the 0.025 and 0.975 quantiles of the corresponding quantities calculated from those bootstrap resamplings.

From a model-building point of view, an interesting inference procedure is to determine if a dual predictor model provides any significant advantage over a single compositional predictor model from Fiksel, Zeger and Datta (2022). There are two related, but not equivalent, scenarios where one would prefer the single-predictor model. The first scenario is where $\theta = 1$, and the proposed model (9) exactly reduces to the single-predictor model from Fiksel, Zeger and Datta (2022) with $\mathbf{X}_1$ as a predictor; similarly, if $\theta = 0$, a similar reduction happens, but $\mathbf{X}_2$ is kept instead. The second scenario is when all of the rows of $\mathbf{B}_2$ (or similarly, $\mathbf{B}_1$) are equal. In the second scenario, any change in the composition of $\mathbf{X}_2$ (or $\mathbf{X}_1$, if $\mathbf{B}_1$ has equal rows) would have no effect on the expected response. In either scenario, it will be true that $E[\mathbf{y} \mid \mathbf{x}_1, \mathbf{x}_2] = E[\mathbf{y} \mid \mathbf{x}_1]$. However, in the second scenario, the model does not exactly reduce to the single-predictor model. To see this, consider the latent variable model where $\mathbf{B}_2$ has equal rows. Then

$$E[\mathbf{y} \mid \mathbf{x}_1, \mathbf{x}_2] = \theta \mathbf{B}_1^T \mathbf{x}_1 + (1 - \theta) \mathbf{B}_2^T \mathbf{x}_2$$

$$
\begin{aligned}
&= \theta \mathbf{B}_1^T \mathbf{x}_1 + (1-\theta)(b_1 \mathbf{1}_{D_{s_2}}, \dots, b_{D_r} \mathbf{1}_{D_{s_2}})^T \mathbf{x}_2 \\
&= \theta \mathbf{B}_1^T \mathbf{x}_1 + (1-\theta)(b_1, \dots, b_{D_r})^T.
\end{aligned}
$$

Above, $b_1, \dots, b_{D_r}$ are non-negative real numbers that sum to 1, and $\mathbf{1}_{D_{s_2}}$ is a D_{s_2}-dimensional all-one vector. Therefore, even if $\mathbf{B}_2$ has all equal rows, the expected value is not necessarily the same as it would be from Fiksel, Zeger and Datta (2022) due to the presence of θ.

Suppose we wished to test $H_0 : \theta = 1$. If H_0 is true, then model (9) reduces to the single predictor model from Fiksel, Zeger and Datta (2022), using only $\mathbf{x}_1$ as a predictor. We propose a bootstrapping procedure to test this hypothesis, which is summarized in Algorithm 3. To test $H_0 : \mathbf{B}_2$ has all equal rows, we propose a permutation test; as this test is similar to that of Fiksel, Zeger and Datta (2022), it is included in the supplementary materials (Rios, Xue and Zhan, 2024).

Before explaining Algorithm 3, it is necessary to review some concepts of the Aitchison geometry (Pawlowsky-Glahn and Egozcue, 2006), where compositional vectors belong to a vector space. Let **a** and **b** be compositional vectors in the interior of $\mathcal{S}^D$. The perturbation operator, which is analogous to vector addition in Euclidean space, is defined as $\mathbf{a} \oplus \mathbf{b} = (a_1 b_1, \dots, a_D b_D)/(\sum_{i=1}^D a_i b_i)$. This operation performs element-wise multiplication on the proportions in **a** and **b**, and then re-normalizes the result. This can also be thought of as applying the composition of **b** to that of **a**. To undo this operation, one can use the inverse perturbation operator, which is defined as $\mathbf{a} \ominus \mathbf{b} = (a_1/b_1, \dots, a_D/b_D)/(\sum_{i=1}^D a_i/b_i)$. It follows that $\mathbf{a} \oplus \mathbf{b} \ominus \mathbf{b} = \mathbf{a}$, so this operation is analogous to subtraction in Euclidean space. The inverse perturbation operator is used in Algorithm 3.

Algorithm 3: Bootstrap Hypothesis Test for $H_0 : \theta = 1$

```
Inputs: Data $\mathbf{X}_1 = \{\mathbf{x}_{1i}, i = 1, \dots, N\}$, $\mathbf{X}_2 = \{\mathbf{x}_{2i}, i = 1, \dots, N\}$, and
  $\mathbf{Y} = \{\mathbf{y}_i, i = 1, \dots, N\}$, number of bootstrap samples $B$
1. Let $\hat{\mathbf{B}}_0$ be the estimate of $\mathbf{B}_1$ under $H_0$ from Fiksel, Zeger and Datta (2022), and let
  $\hat{\mathbf{B}}_1, \hat{\mathbf{B}}_2$ be the estimates of $\mathbf{B}_1, \mathbf{B}_2$ from Algorithm 1.
2. Find $t_0 = \text{KLD}(\mathbf{Y}||E[\mathbf{Y} \mid \mathbf{X}_1, \hat{\mathbf{B}}_0]) - \text{KLD}(\mathbf{Y}||E[\mathbf{Y}, \mathbf{X}_1, \mathbf{X}_2, \hat{\mathbf{B}}_1, \hat{\mathbf{B}}_2])$ using the
  observed data.
3. Using the single-predictor model from Fiksel, Zeger and Datta (2022), regress $\mathbf{Y}$
  on $\mathbf{X}_2$, and store the fitted values as $\hat{\mathbf{Y}}_2 = \{\hat{\mathbf{y}}_{21}, \dots, \hat{\mathbf{y}}_{2N}\}$.
4. Let $\mathbf{y}_{0i} = \mathbf{y}_i \ominus \hat{\mathbf{y}}_{2i},$ for $i = 1, \dots, N$, and let $\mathbf{Y}_0 = \{\mathbf{y}_{01}, \dots, \mathbf{y}_{0N}\}$.
for $i = 1, \dots, B$ do
    5. Generate a bootstrap sample from the predictors by resampling $N$ rows from
      $(\mathbf{X}_1, \mathbf{X}_2, \mathbf{Y}_0)$ with replacement. Denote the resampled values as $\tilde{\mathbf{X}}_1, \tilde{\mathbf{X}}_2$, $\tilde{\mathbf{Y}}$.
    6. Find $\tilde{\mathbf{B}}_0$ by fitting the single-predictor model to $\tilde{\mathbf{X}}_1, \tilde{\mathbf{Y}}$.
    7. Find $\tilde{\mathbf{B}}_1, \tilde{\mathbf{B}}_2$ using Algorithm 1 on $\tilde{\mathbf{Y}}, \tilde{\mathbf{X}}_1, \tilde{\mathbf{X}}_2$.
    8. Find $t_i^* = \text{KLD}(\tilde{\mathbf{Y}}||E[\tilde{\mathbf{Y}} \mid \tilde{\mathbf{X}}_1, \tilde{\mathbf{B}}_0]) - \text{KLD}(\mathbf{Y}||E[\tilde{\mathbf{Y}} \mid \tilde{\mathbf{X}}_1, \tilde{\mathbf{X}}_2, \tilde{\mathbf{B}}_1, \tilde{\mathbf{B}}_2])$.
end
9. p-value $= (1 + \sum_{i=1}^B I(t_i^* \geq t_0))/(B+1)$
Output The p-value.
```

Algorithm 3 takes the compositional predictors and response as input, in addition to a number of bootstrap samples B. If $H_0 : \theta = 1$ is true, then the true model is the single-predictor model with $\mathbf{X}_1$ from Fiksel, Zeger and Datta (2022). In Step 1, Algorithm 3 fits the single and dual-predictor models to the observed dataset. In Step 2, the difference between the KLDs of these models (single-mixture) is computed as a test statistic from the sample. In Steps 3 and 4, the single-predictor model is fit using $\mathbf{X}_2$ as a predictor, and the fitted values

$\hat{\mathbf{Y}}_2$ are stored. In Step 4, the values $\mathbf{y}_{0i} = \mathbf{y}_i \ominus \hat{\mathbf{y}}_{2i}$ are found for each $i = 1, \ldots, N$. This is done so that samples drawn from $(\mathbf{X}_1, \mathbf{X}_2, \mathbf{Y}_0)$ mimic those drawn from the null distribution where $\mathbf{Y}$ only depends on $\mathbf{X}_1$. If θ is close to 1, then $\mathbf{X}_2$ provides little to no information about $\mathbf{Y}$, so $\mathbf{Y}_0$ will not depend on $\mathbf{X}_2$. On the other hand, if θ is close to 0, then $\hat{\mathbf{y}}_{2i} \approx \mathbf{y}_i$, so $\hat{\mathbf{y}}_{0i}$ will be close to $(1/D_r, \ldots, 1/D_r)$, and therefore have little dependence on the value of $\mathbf{x}_{2i}$ for each $i = 1, \ldots, N$. Then, in Steps 5-8, bootstrap samples are taken from $(\mathbf{X}_1, \mathbf{X}_2, \mathbf{Y}_0)$ and the bootstrap statistics t_i^* are computed for each of these bootstrap samples. The p-value that is returned by Algorithm 3 is the proportion of the bootstrapped differences in KLDs that are greater than or equal to the observed difference in KLDs, with 1 added to both the numerator and denominator include the original sample.

From an inferential viewpoint, it is also of interest to quantify the uncertainty of model predictions. Suppose we are given new compositional predictors $\mathbf{x}_{1,N+1}, \mathbf{x}_{2,N+1}$, and let $\mathbf{y}_{N+1}$ be the true but unobserved corresponding compositional response. Our goal is to construct a confidence region C_α such that

$$P(\mathbf{y}_{N+1} \in C_\alpha) \geq 1 - \alpha. \tag{21}$$

In the case of the HTL dataset, this would provide prediction regions for the oil composition of a new biomass given its chemical and elemental compositions, which is quite useful in practice. Since Model (9) does not specify an exact distribution for the compositional response, it is easier to conduct the prediction inference in a nonparametric manner. In particular, we use the conformal inference framework (Shafer and Vovk, 2008; Lei et al., 2018) to provide nonparametric prediction regions for a new observation. Specifically, split conformal inference will be used to efficiently construct prediction intervals via data splitting (Lei et al., 2018). Conformal inference is not unique to compositional data analysis. It is a flexible, nonparametric technique that can be used to construct prediction regions for a variety of regression models for real-valued *i.i.d.* data (Lei et al., 2018) or dependent data (Chernozhukov, Wüthrich and Yinchu, 2018; Yu, Yao and Xue, 2022). Moreover, in the scope of composition-on-composition regression, this framework can be used to form prediction regions for any model with a compositional outcome; this includes not only the proposed model and that of Fiksel, Zeger and Datta (2022), but also the transformation-based models mentioned in Section 2.

In split conformal prediction, the dataset is split into a proper training set $\{(\mathbf{y}_i, \mathbf{x}_{1i}, \mathbf{x}_{2i}) \mid i \in \mathcal{I}_1\}$ of size n_1 and a calibration set $\{(\mathbf{y}_i, \mathbf{x}_{1i}, \mathbf{x}_{2i}) \mid i \in \mathcal{I}_2\}$ of size n_2 such that $n_1 + n_2 = N$ and $\mathcal{I}_1 \cap \mathcal{I}_2 = \emptyset$. Model (9) is fit using the proper training set. Let $\hat{\mathbf{y}}_i$ be the predicted composition for $\mathbf{x}_{1i}, \mathbf{x}_{2i}$ obtained from this fitted model for each $i \in \mathcal{I}_2$ in the calibration set. To determine if a candidate $\mathbf{y}_{\text{cand}}$ is inside C_α, conformity scores are calculated using the KLD. For ease of notation, let $\mathbf{y}_{\text{cand}} = \mathbf{y}_{n_2+1}$ and $\mathcal{I}_2 = \{1, 2, \ldots, n_2\}$. Then define

$$R_{y,i} = \text{KLD}(\mathbf{y}_i || \hat{\mathbf{y}}_i), \quad i = 1, 2, \ldots, n_2 + 1 \tag{22}$$

$$\pi(\mathbf{y}_{n_2+1}) = \frac{1}{n_2 + 1} \sum_{i=1}^{n_2+1} I(R_{y,i} \leq R_{y,n_2+1}) \tag{23}$$

Under the assumption of exchangeability, Lei et al. (2018) show that a $100(1-\alpha)\%$ prediction region for $\mathbf{y}_{n_2+1}$ can then be defined as

$$C_\alpha = \{\mathbf{y}_{\text{cand}} \in \mathcal{S}^{D_r} \mid (n_2 + 1)\pi(\mathbf{y}_{\text{cand}}) \leq \lceil (1 - \alpha)(n_2 + 1) \rceil\} \tag{24}$$

where where the ceiling function $\lceil x \rceil$ maps x to the least integer greater than or equal to x. To construct a prediction region in (24), one would need to repeat the steps in (22) and (23) for many different candidates $\mathbf{y}_{\text{cand}} \in \mathcal{S}^{D_r}$. To illustrate the prediction region, one can test a grid of candidate values over $\mathcal{S}^{D_r}$ and label the points that are in the prediction region. An example of this is shown with the HTL dataset in Section 5.

4. Simulation Studies. There were two main goals to the simulation studies performed in this section. Since log transformation models are the current popular choice for regression with multiple compositional predictors, the first goal was to compare the fit of the proposed transformation-free dual-predictor model (9) to a model that relies on log transformations. The second goal was to study the effect that θ had on the model fit, and to more closely examine the difference in fits between dual-predictor and single-predictor models for different choices of θ. For example, if θ is quite small, how would the fit of the proposed dual predictor model compare to a model that only uses a single predictor? As a consequence, the fit of the proposed dual predictor model was compared to the single-predictor transformation-free models from Fiksel, Zeger and Datta (2022).

To address the above goals, the proposed model was compared with single-predictor direct regression models from Fiksel, Zeger and Datta (2022), as well as an ALR regression model of the form

$$E[\mathrm{alr}(\mathbf{y})_\ell \mid \mathbf{x}_1, \mathbf{x}_2, \boldsymbol{\beta}] = \beta_{0\ell} + \sum_{j=1}^{D_{s_1}-1} \beta_{1j\ell}\mathrm{alr}(\mathbf{x}_1)_j + \sum_{j=1}^{D_{s_2}-1} \beta_{2j\ell}\mathrm{alr}(\mathbf{x}_2)_j \tag{25}$$

for $\ell = 1, \ldots, D_r - 1$. It is important to note that in Model (25), the regression coefficients $\boldsymbol{\beta}$ are not compositional, and they may take any value in the real space. This is different from the family of transformation-free models with compositional regression parameters. As a reminder, the ALR transformation to the predictors and response compositions is described in Section 2. The proposed model was fit via Algorithm 1, which used the grid {0.01, 0.02, . . . , 0.98, 0.99} to search for values of θ.

In order to study the role of θ in the simulations, the responses $\mathbf{y}_i, i = 1, \ldots, N$ were simulated as mixtures of two conditional distributions for (a) $\theta = 0.5$, (b) $\theta = 0.3$, and (c) $\theta = 0.1$. The choice of $\theta = 0.5$ means that both compositional predictors $\mathbf{x}_1$ and $\mathbf{x}_2$ are equally important; for $\theta = 0.1$, $\mathbf{x}_2$ has much more weight than $\mathbf{x}_1$. The middle case of $\theta = 0.3$ was chosen as the midpoint between (a) and (c). For each value of θ, three true models were used to simulate the responses as a mixture of two conditional distributions. These three scenarios are summarized in Table 2, for convenience. This leads to nine total cases.

Scenario	True Coefficients	True Model for $\mathbf{y} \mid \mathbf{x}_1, \mathbf{x}_2$
1	$B_{1j}, B_{2k} \sim$ Dirichlet$(1, \ldots, 1)$ (a) $\theta = 0.5$ (b) $\theta = 0.3$ (c) $\theta = 0.1$	$\mathbf{y}_i \sim \theta$Dirichlet$(\mathbf{B}_1^T \mathbf{x}_{1i}) + (1-\theta)$Dirichlet$(\mathbf{B}_2^T \mathbf{x}_{2i})$
2	$B_{1j}, B_{2k} \sim$ Dirichlet$(1, \ldots, 1)$ (a) $\theta = 0.5$ (b) $\theta = 0.3$ (c) $\theta = 0.1$	$\mathbf{y}_i \sim \mathrm{alr}^{-1}(N(\mathrm{alr}(\theta \mathbf{B}_1^T \mathbf{x}_{1i} + (1-\theta)\mathbf{B}_2^T \mathbf{x}_{2i}), 1)$
3	$\beta_{1j\ell} \sim N(1, 0.5), \beta_{2j\ell} \sim N(1, 0.5)$ (a) $\theta = 0.5$ (b) $\theta = 0.3$ (c) $\theta = 0.1$	$\mathbf{y}_i \sim \theta\mathrm{alr}^{-1}(N(E[\mathrm{alr}(y \mid \mathbf{x}_{1i}, \boldsymbol{\beta}_1)], 1))$ $+(1-\theta)\mathrm{alr}^{-1}(N(E[\mathrm{alr}(y \mid \mathbf{x}_{2i}, \boldsymbol{\beta}_2)], 1))$

TABLE 2
Simulation Settings

Scenarios 1 and 2 in Table 2 both simulated the responses as a mixture of distributions that were parameterized by known transition matrices $\mathbf{B}_1$ and $\mathbf{B}_2$, so it was expected that the proposed methodology should be an appropriate choice in these cases for $\theta = 0.5, 0.3$. In Scenario 3, the responses were simulated as a mixture of single-predictor ALR models with known regression vectors $\boldsymbol{\beta}_1$ and $\boldsymbol{\beta}_2$ that are not compositional. In this scenario, since the underlying mechanism was a mixture of ALR models, it was expected that the dual-predictor ALR model (25) will perform well in terms of model fit. The main reason why we did not directly simulate the data using the ALR model (25) was so that we could also study the role of θ in Scenario 3. Moreover, it was interesting to examine how well the ALR model performed when the true response was a mixture of conditional distributions, as when one fits the ALR model (25), they are not inherently assuming the response is a mixture distribution.

In each scenario in Table 2, the predictors $\mathbf{x}_1, \mathbf{x}_2$ were independent Dirichlet$(1, \dots, 1)$ random variables of dimension $D_s = D_{s_1} = D_{s_2}$. Each scenario was executed for sample sizes $N = 100, 200, \dots, 700$, with dimensions $D_s = D_r = 3$. For each N, the log mean KLD was used to compare the fit of the models. The log mean KLD was taken over an independently generated test set of size 10000. The log mean KLDs are displayed in Figure 1, and a table of these values is provided in Section S5 of the supplementary materials (Rios, Xue and Zhan, 2024).

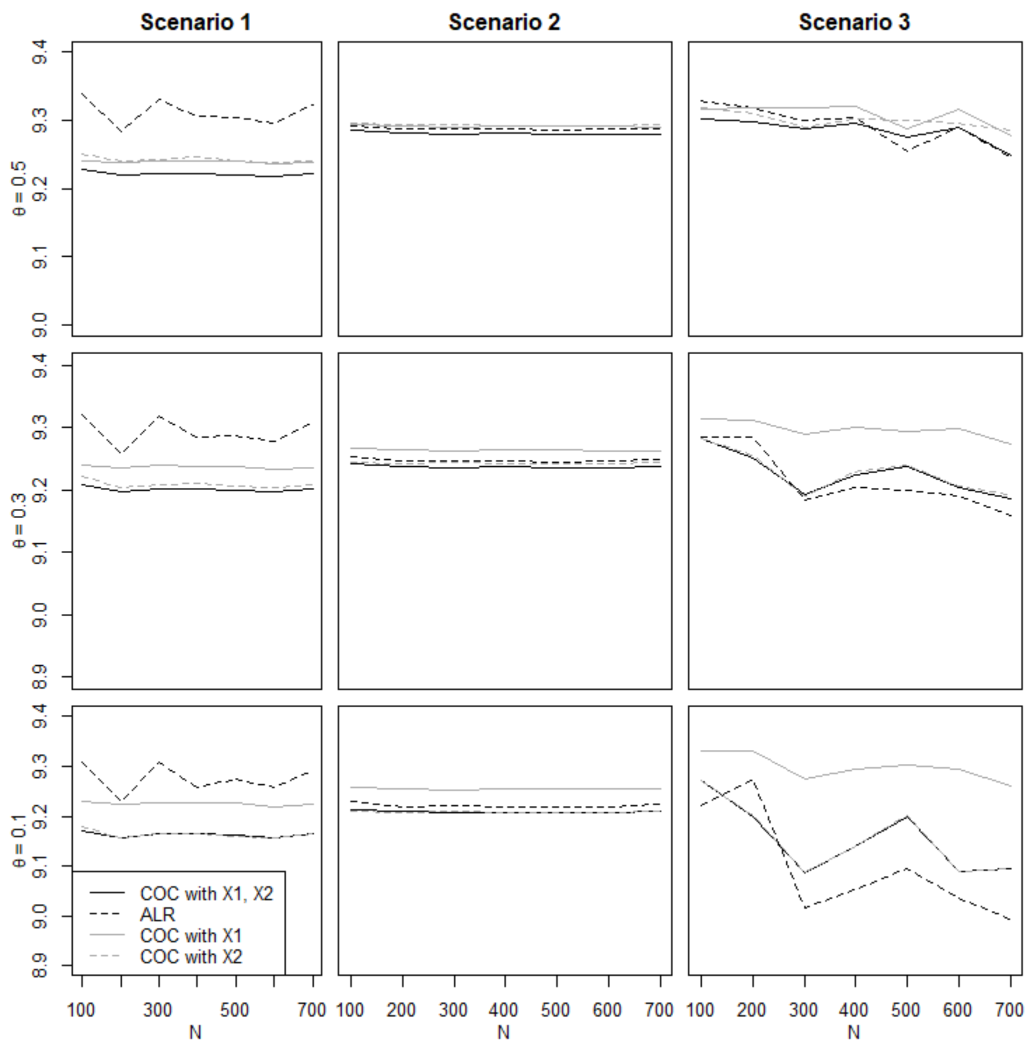

FIG 1. *Log Mean KLD Comparison for Three Models for Scenarios 1(a,b,c), 2(a,b,c), and 3(a,b,c)*

As shown in Figure 1, the proposed latent variable composition-on-composition regression model (solid line) had the lowest average KLD in Scenarios 1 and 2 for $\theta = 0.5, 0.3$. This was expected because in both of these scenarios, θ was not too close to 0 or 1, and the regression model was taken to be a mixture of two distributions that are parameterized by Markov transition matrices, which was the core assumption behind the proposed model (9).

Scenario	**N**						
	100	200	300	400	500	600	700
1(a), $\theta = 0.5$	0.53	0.50	0.50	0.50	0.50	0.50	0.50
2(a), $\theta = 0.5$	0.50	0.50	0.50	0.50	0.50	0.50	0.50
3(a), $\theta = 0.5$	0.50	0.50	0.50	0.50	0.50	0.39	0.49
1(b), $\theta = 0.3$	0.35	0.30	0.30	0.30	0.30	0.30	0.30
2(b), $\theta = 0.3$	0.30	0.30	0.30	0.30	0.30	0.30	0.30
3(b), $\theta = 0.3$	0.28	0.30	0.18	0.19	0.23	0.11	0.18
1(c), $\theta = 0.1$	0.17	0.10	0.10	0.16	0.10	0.10	0.10
2(c), $\theta = 0.1$	0.19	0.10	0.10	0.10	0.10	0.12	0.10
3(c), $\theta = 0.1$	0.08	0.10	0.01	0.03	0.03	0.01	0.01

TABLE 3
Estimates of θ for Each Scenario

In all scenarios, when $\theta = 0.1$, the performance between the proposed dual predictor model and the COC model with $\mathbf{x}_2$ only was nearly identical in terms of log mean KLD. This is shown in Figure 1 by the overlapping gray and blue lines in the final row.

In the third scenario, the true regression model was a mixture of two ALR models (25). In Scenario 3(a), where $\theta = 0.5$, the proposed dual predictor model performed better than ALR for $N \leq 400$, and had similar performance to the ALR model for $N \geq 600$. In Scenarios 3(b) and (c) where $\theta = 0.3$ and $\theta = 0.1$, respectively, the proposed model performed better than ALR in terms of average log KLD for $N \leq 300$. The ALR model performed the best in Scenario 3(c), but this was because, with $\theta = 0.1$, the responses were mainly following an ALR model based on $\mathbf{x}_2$ only.

Overall, it was demonstrated that when the underlying regression model was a mixture of distributions, the proposed model appeared to either have the lowest KLD or was reasonably close in terms of KLD to the best alternative. In all scenarios where $\theta = 0.5$, it was much better to use the proposed model than the direct regression model from Fiksel, Zeger and Datta (2022), which only included one predictor. When $\theta = 0.1$, more emphasis was placed on the $\mathbf{x}_2$ predictor as opposed to $\mathbf{x}_1$. In these cases, the direct regression model for $\mathbf{x}_2$ had lower mean KLD than the direct regression model for $\mathbf{x}_1$.

The estimates of θ for each scenario and each value of $N = 100, \ldots, 700$ are displayed in Table 3. As shown in Table 3, the estimated values of θ were quite close to the true values of θ for scenarios 1 and 2, especially for large N. In scenarios 3(b) and 3(c), the value of θ tended to be underestimated as N increased. This was likely because the true model used to generate the responses did not use compositional transition matrices, which created bias in the parameter estimates. Similar comparisons were done for the case when $p = 3$ in Section S5 of the supplementary materials.

It was also of interest to briefly examine the Type I error rates and power of the hypothesis tests proposed in Section 3.4. We considered two hypotheses of interest. The first was $H_0 : \theta = 0$. We also considered $H_0 : \mathbf{B}_1$ has equal rows; simulated Type I error rates and power for testing this hypothesis can be found in the supplementary materials (Rios, Xue and Zhan, 2024). We set $\alpha = 0.05$. To examine Type I error, data of size $N = 100$ were simulated 1000 times according to Scenarios 1,2, and 3 for $\theta = 0$. For testing Type I error and power, Scenario 1 was modified. The responses were simulated as $\mathbf{y}_i \sim \theta\text{Dirichlet}(2\mathbf{B}_1^T\mathbf{x}_{1i}) + (1 - \theta)\text{Dirichlet}(2\mathbf{B}_2^T\mathbf{x}_{2i})$, as increasing the concentration of the Dirichlet parameters resulted in higher power and more stable estimates. The bootstrap test used $B = 100$. The simulated Type I error was then recorded as the proportion of times that H_0 was rejected. In Table 4, the Type I error rates are shown. All of the rates were close to the nominal Type I error rate of $\alpha = 0.05$, except for the Dirichlet scenario, which was lower.

To examine the power of the bootstrap test, data of size $N = 500$ were simulated 100 times according to Scenarios 1,2, and 3 for $\theta = 0.5, 0.75$, $B = 500$, and $\alpha = 0.05$. The power was

Scenario	1	2	3
Type I Error	0.024	0.058	0.049

TABLE 4

Simulated Type I Error Rates for Bootstrap Test of $H_0 : \theta = 0$, $N = 100$, 1000 simulations

estimated by the proportion of times H_0 was rejected. As shown in Table 5, as θ increased, the power increased. The power was the highest in Scenario 3, when the data were generated by the ALR mechanism and lowest in Scenario 2. When the effect size was largest ($\theta = 0.75$), the power was over 95% for all scenarios. We note that larger sample sizes and increased numbers of permutations are very helpful for achieving higher power.

	θ	
Scenario	0.5	0.75
1	0.84	1.00
2	0.61	0.97
3	1.00	1.00

TABLE 5

Simulated Power for Bootstrap Test of $H_0 : \theta = 0$.

Simulations were also used to examine the coverage rates of the split conformal prediction regions for a new composition at $\mathbf{x}_{1,N+1} = \mathbf{x}_{2,N+1} = (1/3, 1/3, 1/3)$ and for $\mathbf{x}_{1,N+1} = \mathbf{x}_{2,N+1} = (0.5, 0.25, 0.25)$ For each scenario, the true response $\mathbf{y}_{N+1}$ was drawn from the corresponding distribution of the response, shown in Table 2. The response data (with $N = 500$ data points) were simulated 100 times. Each time, the simulated dataset was split equally into training and testing data. The proposed model was fit to the training dataset, and the testing dataset was used to determine if the true value of $\mathbf{y}_{N+1}$ was contained in the 95% conformal prediction region. The results are summarized in Table 6, which shows the empirical coverage of the 95% conformal prediction regions at the center point of $\mathbf{x}_1 = \mathbf{x}_2 = (1/3, 1/3, 1/3)$ and $\mathbf{x}_{1,N+1} = \mathbf{x}_{2,N+1} = (0.5, 0.25, 0.25)$ for $N = 500$. We can see that the empirical coverage fluctuated about 0.95 across the scenarios. In general, for Scenarios 1 and 2, the confidence regions captured the true simulated response $\mathbf{y}_{N+1}$ slightly more than 95% of the time. In Scenario 3(b), the true response is captured slightly less than 95% of the time.

	Scenario								
$\mathbf{x}_{1,N+1} = \mathbf{x}_{2,N+1}$	1(a)	1(b)	1(c)	2(a)	2(b)	2(c)	3(a)	3(b)	3(c)
(1/3, 1/3, 1/3)	0.96	0.98	0.99	0.97	0.98	0.99	0.99	0.94	0.98
(0.5, 0.25, 0.25)	0.95	0.99	0.99	0.97	0.99	0.99	0.99	0.99	0.99

TABLE 6

Estimated Coverage Rates for 95% Prediction Region for $\mathbf{y}_{N+1}$ with $\mathbf{x}_{1,N+1} = \mathbf{x}_{2,N+1}$

5. Application to HTL Data. To show the efficacy of the proposed methodology, the dual compositional predictor model was fit to the HTL dataset (Subramanya et al., 2023). Hydrothermal liquefaction (HTL) of biomass occurs at high temperatures and pressures. Under these conditions, the biomass undergoes reactions that produce several components in bio-oil. The response of interest was the oil composition ($\mathbf{y}$) which had nine components $y_1, \ldots, y_9$. There were two compositional predictors of interest: the chemical composition of the biomass ($\mathbf{x}_1$) and the elemental composition of the biomass ($\mathbf{x}_2$). These predictors had four components each. The response and predictor variables are summarized in Table 7. The data contained $N = 413$ complete-case rows. The full HTL data set is available on Mendeley data (Mahadevan et al., 2023).

Variable	**Description**
y_1	Esters
y_2	Oxygenated single ring aromatics
y_3	Furans
y_4	Long chain fatty acids
y_5	Long chain alcohols
y_6	Aldehydes and Ketones
y_7	N-containing compounds
y_8	Aliphatics
y_9	Polycyclic aromatics
x_{11}	Carbohydrate
x_{12}	Protein
x_{13}	Lipid
x_{14}	Lignin
x_{21}	Carbon (C)
x_{22}	Hydrogen (H)
x_{23}	Nitrogen (N)
x_{24}	Oxygen (O)

TABLE 7

Response and Predictor Variables in HTL Dataset

We first fit the proposed dual model to the full dataset. For the dual predictor model, estimates of $\mathbf{B}_1, \mathbf{B}_2$, and θ were found using Algorithm (1). The grid points for Algorithm (1) were $(0.01, 0.02, \ldots, 0.99)$. The KLD for the dual predictor model is minimized at $\hat{\theta} = 0.76$. The value of $\hat{\theta} = 0.76$ indicates that an estimated 76% of the biomasses in the HTL dataset belonged to a latent group whose oil composition is better characterized by their biochemical composition, as opposed to their elemental composition. Algorithm 1 converged in 11 iterations. The corresponding estimates $\hat{B}_1, \hat{B}_2$ are displayed in Table 8 and Table 9, respectively.

Chemical	y_1	y_2	y_3	y_4	y_5	y_6	y_7	y_8	y_9
Carbohydrate	0.31	0.23	0.07	0.08	0	0.18	0.01	0.12	0
Protein	0.14	0	0	0.06	0.09	0	0.68	0.03	0
Lipid	0.05	0.04	0	0.49	0	0	0	0.21	0.21
Lignin	0	0.85	0.02	0	0	0.05	0	0	0.08

TABLE 8

Estimates for $\boldsymbol{B}_1$ in Model (9), Chemical Composition

Element	y_1	y_2	y_3	y_4	y_5	y_6	y_7	y_8	y_9
Carbon	0.09	0.24	0	0	0	0	0.06	0	0.61
Hydrogen	0	0	0	1	0	0	0	0	0
Nitrogen	0	0	0	0	0	0	1	0	0
Oxygen	0	0	0	0.47	0.27	0.26	0	0	0

TABLE 9

Estimates for $\boldsymbol{B}_2$ in Model (9), Elemental Composition

Chemical	y_1	y_2	y_3	y_4	y_5	y_6	y_7	y_8	y_9
Carbohydrate	0.25	0.19	0.05	0.13	0.04	0.17	0.02	0.09	0.06
Protein	0.13	0.01	0	0.10	0.10	0.01	0.56	0.03	0.06
Lipid	0.05	0.08	0	0.42	0	0	0.02	0.16	0.27
Lignin	0	0.71	0.01	0.06	0.01	0.07	0	0	0.14

TABLE 10

Estimates for $\boldsymbol{B}_1$ in Chemical Only Model

Element	y_1	y_2	y_3	y_4	y_5	y_6	y_7	y_8	y_9
Carbon	0.29	0.09	0	0.08	0.02	0	0.20	0.11	0.21
Hydrogen	0	0	0	1	0	0	0	0	0
Nitrogen	0	0	0	0	0	0	1	0	0
Oxygen	0.06	0.41	0.06	0.09	0.09	0.24	0	0.05	0

TABLE 11

Estimates for $\boldsymbol{B}_2$ in Elemental Only Model

The interpretation of $\hat{B}_1$ and $\hat{B}_2$, is straightforward. To interpret the coefficients in Tables 8 and 9, recall that if component i of the chemical composition of the biomass is increased by a small Δ at the cost of decreasing another distinct component $j \neq i$ of the chemical composition by Δ, then the expected change in the oil composition is $\hat{\theta}\Delta(\hat{B}_{1i*} - \hat{B}_{1j*})$, assuming that the elemental composition of the substance is unchanged. For example, fixing the elemental composition, if we increase Protein by Δ and decrease Lipids by Δ, then the expected changes in y_1 (Esters) and y_2 (Oxygenated single ring aromatics) are 0.068Δ and -0.030Δ, respectively, after rounding to 2 decimal places. Similarly, if the chemical composition of a biomass is fixed, then increasing Nitrogen by Δ and decreasing Oxygen by Δ yields an expected increase of $(1-\hat{\theta})(1-0)\Delta = 0.24\Delta$ in y_7 (N-containing compounds) and an expected decrease of 0.06Δ in y_6 (Aldehydes and Ketones). It is also intuitive that the third row of $\hat{B}_2$ is concentrated entirely on y_7, as y_7 corresponds to Nitrogen-containing compounds, and the third elemental predictor is Nitrogen.

We considered three models for comparison. The first was the proposed model (9) that used both the chemical and elemental compositions as predictors. The second model only used the chemical composition as a predictor. The third model only used the elemental composition as a predictor. The second and third models were fit using the R package codalm, which is based on the methodology from Fiksel, Zeger and Datta (2022).

To compare the dual predictor model with a model that only used chemical composition, we first tested $H_0 : \theta = 1$ and $H_0 : \theta = 0$ using Algorithm 3 with $B = 500$. The test for $H_0 : \theta = 1$ had a p-value of 0.0019, and the test for $H_0 : \theta = 0$ had a p-value of 0.0159. Both tests had p-values less than 0.05, which yielded sufficient evidence to reject H_0 in each case. We also tested $H_0 : \mathbf{B}_2$ has equal rows versus the alternative that H_0 is not true. This can be done using a global permutation test, as described in the supplementary materials (Rios, Xue and Zhan, 2024). 500 permutations were used to shuffle the rows of $\mathbf{x}_2$ for fixed $\mathbf{x}_1, y$, and the difference in KLDs (reduced - full) was calculated for each permutation. The observed difference in KLDs was 2.242, with a p-value of 0.0059. Since this p-value was low, there was sufficient evidence to reject H_0 in favor of H_A. When testing $H_0 : \mathbf{B}_1$ has equal rows, the observed difference in KLDs was 68.811 with a p-value of 0.0019. We similarly rejected H_0 in this case. For comparison, the values of $\mathbf{B}_1$ under the chemical only model are displayed in Table 10, and the values of $\mathbf{B}_2$ under the elemental only model are displayed in Table 11. By comparing these tables to the estimates for $\mathbf{B}_1, \mathbf{B}_2$ under the dual model, it was apparent that there are several differences in the estimates for $\mathbf{B}_2$. For example, in the elemental-only model, Oxygen had a much stronger effect on y_2 (Oxygenated single ring aromatics) and y_6 (Aldehydes and Ketones) than it did in the dual predictor model.

To further analyze the data, we constructed 95% confidence intervals for $\theta, \mathbf{B}_1$, and $\mathbf{B}_2$ using 500 bootstrap resamples. To create each resample, the original data rows $(\mathbf{x}_{1i}, \mathbf{x}_{2i}, \mathbf{y}_i)$ were resampled with replacement from the original dataset until $N = 413$ points were obtained. For each element of $\mathbf{B}_1$ and $\mathbf{B}_2$, pointwise 95% confidence intervals were found using the 0.025 and 0.975 quantiles of the bootstrap resamples. For each row of $\mathbf{B}_1$ and $\mathbf{B}_2$, a Bonferroni correction was performed by using $\alpha = 0.05/9$ instead to obtain simultaneous confidence intervals that maintain 95% coverage across each row. These intervals are shown in Figures 2 and 3 for $\mathbf{B}_1$ and $\mathbf{B}_2$, respectively. In these figures, the gray confidence intervals represent the pointwise percentile-based 95% bootstrap confidence intervals, and the black lines represent the Bonferroni-adjusted 95% confidence intervals. In all cases, the Bonferroni intervals (black bands) were wider than their pointwise counterparts. The difference between the Bonferroni and pointwise intervals was quite minor for Carbohydrates, Proteins, Lignins, and most of the components of Carbon and Oxygen. Additionally, most components of $\mathbf{B}_1$ and $\mathbf{B}_2$ that were estimated to be zero had confidence limits that were very close to zero, and therefore did not appear in the figures. The exceptions to these were components 1, 4, and 9 of the Hydrogen row in $\mathbf{B}_2$, and components 7 and 9 of the Nitrogen row in $\mathbf{B}_2$, all of which had Bonferroni confidence limits that were quite wide, indicating more uncertainty in these estimates. Finally, the 95% percentile-based confidence interval for θ was $(0.71, 0.85)$.

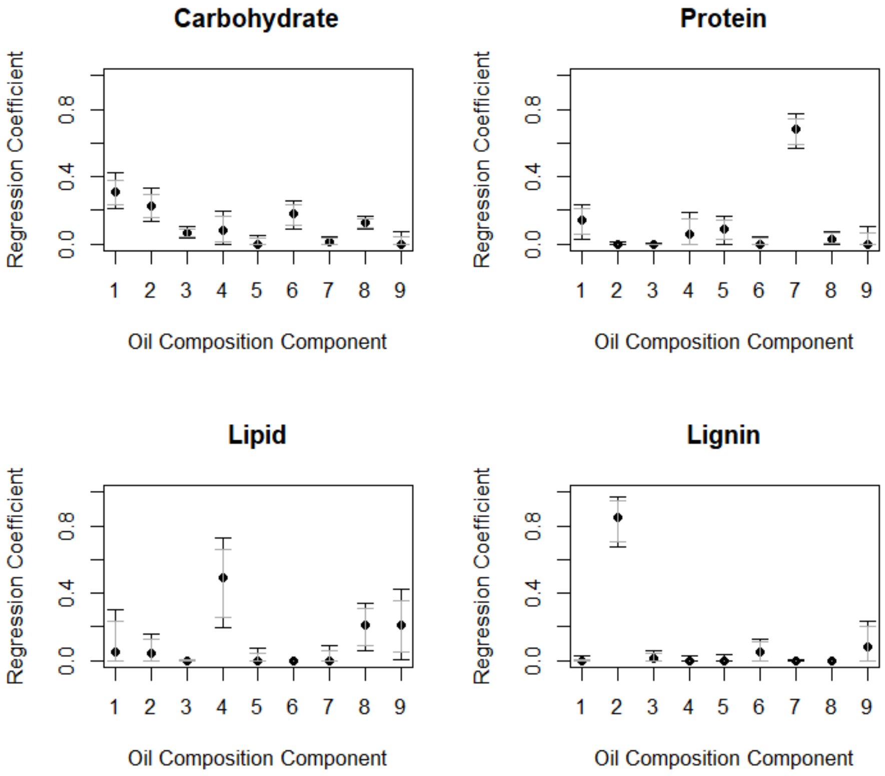

FIG 2. *95% Bootstrap Confidence Intervals for B_1 (regression coefficients for chemical composition). Black lines are Bonferroni adjusted. Gray lines represent pointwise limits.*

Finally, to illustrate the use of split conformal inference, the HTL dataset was randomly separated into two equal partitions for training and testing. The dual predictor model was fit

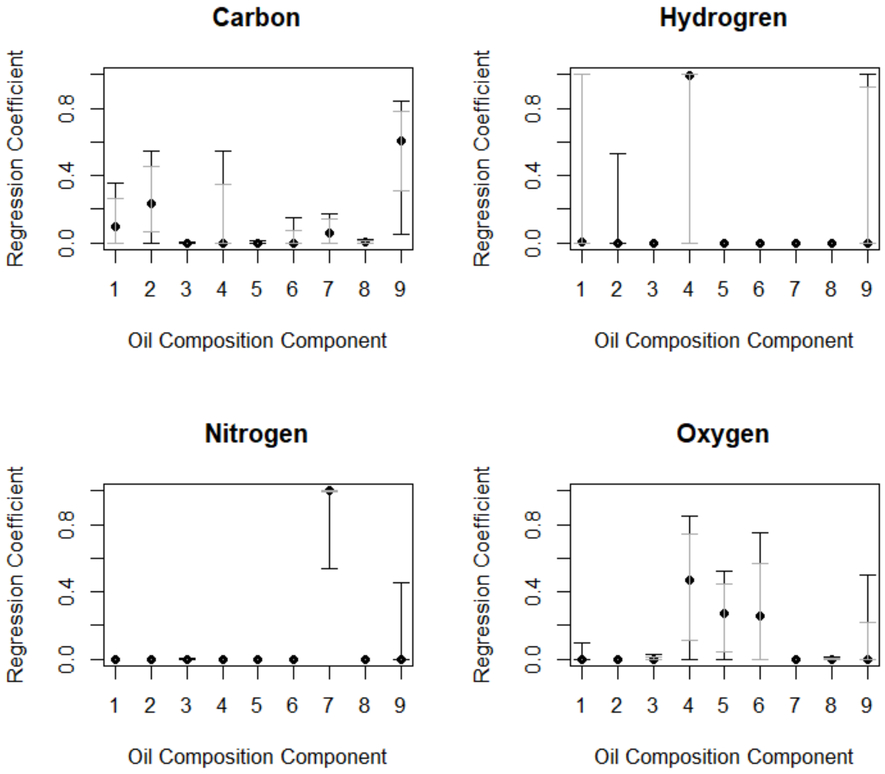

FIG 3. *95% Bootstrap Confidence Intervals for B_2 (regression coefficients for elemental composition). Black lines are Bonferroni adjusted. Gray lines represent pointwise limits.*

to the training dataset. The goal was to form and visualize a 95% prediction region for a new oil composition $\mathbf{y}_{N+1}$ given predictors $\mathbf{x}_1 = \mathbf{x}_2 = (0.25, 0.25, 0.25, 0, 25)$. An equally spaced grid on a 9-dimensional simplex was used to find candidates for $\mathbf{y}_{N+1}$. Each candidate in the grid was labeled as being inside the 95% confidence region or not. Since the confidence region was a 9-dimensional space, it was visualized using $\binom{9}{2} = 36$ bivariate scatterplots, which are shown in Figure 4. In Figure 4, points that are inside the 95% prediction region for $\mathbf{y}_{N+1}$ are colored in gray. Points that are not inside the region are dark. This gives us some insight as to what types of oil composition are implausible to consider for these values of $\mathbf{x}_1$ (chemical composition) and $\mathbf{x}_2$ (elemental composition). For example, by looking at the third row of the figure, it seems that having high values of y_3 (Furans) is unlikely, as most of the dots in the third row are dark for $y_3 > 0.4$.

6. Conclusion. To the best of our knowledge, this is the first paper that proposed a transformation-free compositional latent variable model that can accommodate more than one compositional predictor. The proposed latent variable model represents the expected response as a convex combination of two (or more) conditional expectations. The theoretical results of Fiksel, Zeger and Datta (2022) are extended to prove that the parameter estimates from the EM algorithm minimize the Kullback-Leibler distance between the observed and expected responses. An algorithm is shown for estimating the mixture parameter θ in addition to the model covariates. Furthermore, Section 3.3 describes the inclusion of more than two compositional covariates. Overall, these are significant extensions of the existing methodology proposed in Fiksel, Zeger and Datta (2022).

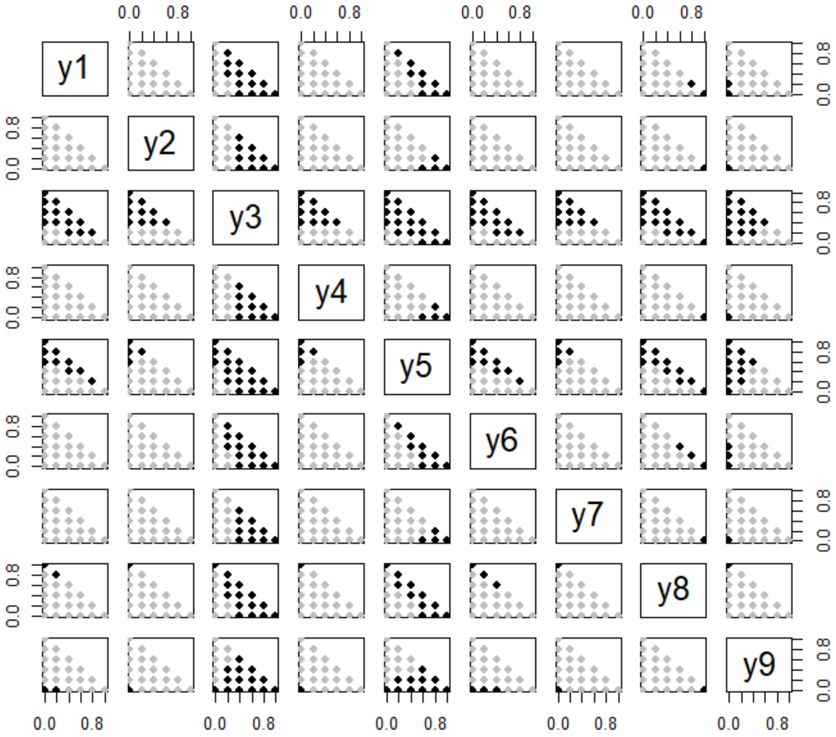

FIG 4. *Two-Dimensional Projections of a 95% Prediction Region for* $\boldsymbol{y}_{N+1}$ *at* $\boldsymbol{x}_{1,N+1} = \boldsymbol{x}_{2,N+1} = (0.25, 0.25, 0.25, 0.25)$. *Gray points are in the prediction region; black points are not.*

This methodology can be applied to compositions of higher dimension, and the resulting parameter estimates are easy to interpret. Furthermore, the proposed model can accommodate zeros in both the responses and covariates, which is an advantage over existing log transformation methods. These advantages are demonstrated in Section 5, where the latent variable dual predictor model is applied to the HTL dataset, which is an important application in chemical recycling. The dimension of the oil composition ($\mathbf{y}$) is much larger than those of the chemical ($\mathbf{x}_1$) and elemental ($\mathbf{x}_2$) compositions. Furthermore, zeros exist in $\mathbf{y}, \mathbf{x}_1$, and $\mathbf{x}_2$. The interpretation of the parameter estimates showed that, among many things, the elemental composition was useful for predicting the percentages of nitrogen-containing compounds and long-chain fatty acids. The majority of the components in the oil are explained by changes in protein, carbohydrates, and other chemical compositions of the biomass. It is our hope that this methodology will be of use to analysts dealing with multiple compositional predictors.

There is much future work to be done in this area. The EM algorithm provides quick and numerically stable solutions for estimating the elements of the transition matrices $\mathbf{B}_1, \mathbf{B}_2$. However, these estimates assume that θ is constant. Allowing for subject-level weights would be an interesting and more flexible extension of this work. These weights could be estimated as functions of compositional covariates, or non-compositional covariates, such as time or temperature. Furthermore, the algorithms used to estimate the mixture parameters in this paper may not scale well computationally for a large number of compositional predictors. More work could be done to find an efficient way to estimate these parameters in an even higher dimensional setting. It would also be interesting to see this methodology applied to more real datasets. Finally, this model estimated parameters by minimizing the KL-distance between the observed and expected responses. There could be other criteria or methods that are useful for parameter estimation in this setting.

Acknowledgments. Nicholas Rios and Lingzhou Xue were supported in part by the National Science Foundation grants (DMS-2210775, CCF-2007823, DMS-1953189) and the National Institute of General Medical Sciences grant (1R01GM152812). Xiang Zhan was supported in part by the National Natural Science Foundation of China (grant no. 12371287). The authors would like to thank the referees, the Associate Editor, and the Editor for constructive comments that improved this article. Xiang Zhan is the corresponding author.

SUPPLEMENTARY MATERIAL

Supplementary Information: Proofs and Additional Simulation Results
This file provides more detailed information on the proof of Corollary 1 and Corollary 2, the permutation test referenced in Section 3, and additional simulation results of interest.

HTL Dataset and Source Code
This zip file contains the HTL dataset and all R code used to produce the analyses shown in this paper and the supplementary information.

REFERENCES

AGARWAL, A. and XUE, L. (2020). Model-based clustering of nonparametric weighted networks with application to water pollution analysis. *Technometrics* **62** 161–172.

AITCHISON, J. (1982). The statistical analysis of compositional data. *Journal of the Royal Statistical Society: Series B (Methodological)* **44** 139–160.

AITCHISON, J. and BACON-SHONE, J. (1984). Log contrast models for experiments with mixtures. *Biometrika* **71** 323–330.

CHEN, J. and LI, H. (2013). Variable selection for sparse Dirichlet-multinomial regression with an application to microbiome data analysis. *The Annals of Applied Statistics* **7** 418–442.

CHEN, J., ZHANG, X. and LI, S. (2017). Multiple linear regression with compositional response and covariates. *Journal of Applied Statistics* **44** 2270–2285.

CHERNOZHUKOV, V., WÜTHRICH, K. and YINCHU, Z. (2018). Exact and robust conformal inference methods for predictive machine learning with dependent data. In *Conference On learning theory* 732–749. PMLR.

DESARBO, W. S., CHEN, Q. and BLANK, A. S. (2017). A parametric constrained segmentation methodology for application in sport marketing. *Customer Needs and Solutions* **4** 37–55.

DOUMA, J. C. and WEEDON, J. T. (2019). Analysing continuous proportions in ecology and evolution: A practical introduction to beta and Dirichlet regression. *Methods in Ecology and Evolution* **10** 1412–1430.

EGOZCUE, J. J., PAWLOWSKY-GLAHN, V., MATEU-FIGUERAS, G. and BARCELO-VIDAL, C. (2003). Isometric logratio transformations for compositional data analysis. *Mathematical Geology* **35** 279–300.

FIKSEL, J., ZEGER, S. and DATTA, A. (2022). A transformation-free linear regression for compositional outcomes and predictors. *Biometrics* **78** 974–987.

GOLLAKOTA, A., KISHORE, N. and GU, S. (2018). A review on hydrothermal liquefaction of biomass. *Renewable and Sustainable Energy Reviews* **81** 1378–1392.

GUIRGUIS, P. M., SESHASAYEE, M. S., MOTAVAF, B. and SAVAGE, P. E. (2024). Review and assessment of models for predicting biocrude yields from hydrothermal liquefaction of biomass. *RSC Sustainability* **2** 736–756.

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

LEE, K. H. and XUE, L. (2018). Nonparametric finite mixture of Gaussian graphical models. *Technometrics* **60** 511–521.

LEE, K. H., CHEN, Q., DESARBO, W. S. and XUE, L. (2022). Estimating finite mixtures of ordinal graphical models. *Psychometrika* **87** 83–106.

LEI, J., G'SELL, M., RINALDO, A., TIBSHIRANI, R. J. and WASSERMAN, L. (2018). Distribution-free predictive inference for regression. *Journal of the American Statistical Association* **113** 1094–1111.

LIN, W., SHI, P., FENG, R. and LI, H. (2014). Variable selection in regression with compositional covariates. *Biometrika* **101** 785–797.

LU, J., LIU, Z., ZHANG, Y. and SAVAGE, P. E. (2018). Synergistic and antagonistic interactions during hydrothermal liquefaction of soybean oil, soy protein, cellulose, xylose, and lignin. *ACS Sustainable Chemistry & Engineering* **6** 14501–14509.

MAHADEVAN, S., RIOS, N., KOLLAR, A. J., STOFANAK, R., MALONEY, K., WALTZ, K. E., RANE, C., ENDLURI, S. and SAVAGE, P. E. (2023). Dataset for oil composition, yield from Hydrothermal Liquefaction of biomass. *Mendeley Data* **1**. https://doi.org/10.17632/s38wv3fvpz.1

MCKENDRY, P. (2002). Energy production from biomass (part 1): overview of biomass. *Bioresource technology* **83** 37–46.

MCLACHLAN, G. J. and KRISHNAN, T. (2007). *The EM Algorithm and Extensions*. John Wiley & Sons.

MOSIMANN, J. E. (1962). On the compound multinomial distribution, the multivariate β-distribution, and correlations among proportions. *Biometrika* **49** 65–82.

PAWLOWSKY-GLAHN, V. and EGOZCUE, J. J. (2006). Compositional data and their analysis: an introduction. *Geological Society, London, Special Publications* **264** 1–10.

PEARSON, K. (1897). On a form of spurious correlation which may arise when indices are useed in the measurement of organs. In *Royal Soc., London, Proc.* **60** 489–502.

RIOS, N., XUE, L. and ZHAN, X. (2024). Supplement to "A Latent Variable Mixture Model for Composition-on-Composition Regresssion with Application to Chemical Recycling". https://doi.org/10.12124/[provided by typesetter]

SHAFER, G. and VOVK, V. (2008). A Tutorial on Conformal Prediction. *Journal of Machine Learning Research* **9** 371–421.

SHAHBEIK, H., PANAHI, H. K. S., DEHHAGHI, M., GUILLEMIN, G. J., FALLAHI, A., HOSSEINZADEH-BANDBAFHA, H., AMIRI, H., REHAN, M., RAIKWAR, D., LATINE, H. et al. (2024). Biomass to biofuels using hydrothermal liquefaction: A comprehensive review. *Renewable and Sustainable Energy Reviews* **189** 113976.

SRINIVASAN, A., XUE, L. and ZHAN, X. (2021). Compositional knockoff filter for high-dimensional regression analysis of microbiome data. *Biometrics* **77** 984–995.

SRINIVASAN, A., XUE, L. and ZHAN, X. (2022). Identification of microbial features in multivariate regression under false discovery rate control. *Computational Statistics & Data Analysis* 107621.

SUBRAMANYA, S. M., RIOS, N., KOLLAR, A., STOFANAK, R., MALONEY, K., WALTZ, K., POWERS, L., RANE, C. and SAVAGE, P. E. (2023). Statistical Models for Predicting Oil Composition from Hydrothermal Liquefaction of Biomass. *Energy & Fuels* **37** 6619–6628.

TANG, Z.-Z. and CHEN, G. (2019). Zero-inflated generalized Dirichlet multinomial regression model for microbiome compositional data analysis. *Biostatistics* **20** 698–713.

VALDEZ, P. J., TOCCO, V. J. and SAVAGE, P. E. (2014). A general kinetic model for the hydrothermal liquefaction of microalgae. *Bioresource Technology* **163** 123–127.

YU, X., YAO, J. and XUE, L. (2022). Nonparametric estimation and conformal inference of the sufficient forecasting with a diverging number of factors. *Journal of Business & Economic Statistics* **40** 342–354.

ZHANG, L., DOU, X., YANG, Z., YANG, X. and GUO, X. (2021). Advance in hydrothermal bio-oil preparation from lignocellulose: effect of raw materials and their tissue structures. *Biomass* **1** 74–93.