Biometrics 000, 1–27 DOI: 000
April 2020

A Transformation-free Linear Regression for Compositional Outcomes and Predictors

Jacob Fiksel*, Scott Zeger, and Abhirup Datta

Department of Biostatistics, Johns Hopkins University, 615 North Wolfe Street, Baltimore, MD 21205, USA

**email:* jfiksel@gmail.com

Summary: Compositional data are common in many fields, both as outcomes and predictor variables. The inventory of models for the case when both the outcome and predictor variables are compositional is limited and the existing models are often difficult to interpret in the compositional space, due to their use of complex log-ratio transformations. We develop a transformation-free linear regression model where the expected value of the compositional outcome is expressed as a single Markov transition from the compositional predictor. Our approach is based on estimating equations thereby not requiring complete specification of data likelihood and is robust to different data generating mechanisms. Our model is simple to interpret, allows for 0s and 1s in both the compositional outcome and covariates, and subsumes several interesting subcases of interest. We also develop permutation tests for linear independence, and equality of effect sizes of two components of the predictor. Finally, we show that despite its simplicity, our model accurately captures the relationship between compositional data using two data sets from education and medical research.

Key words: Compositional data; EM Algorithm; GMM; KLD Loss Function; Transformation-free

1. Introduction

Compositional data, also referred to as fractional data (Mullahy, 2015; Murteira and Ramalho, 2016), consist of vectors constrained to lie in the unit simplex, $\mathbb{S}^D$, where $\mathbb{S}^D = \{(x_1, x_2, \ldots, x_D)^{'}|x_j \geqslant 0, i = j, \ldots, D; \sum_{i=j}^{D} x_j = 1\}$. Compositional data appear in many fields, such as econometrics (Papke and Wooldridge, 1996), geochemistry (Templ et al., 2008), physical activity research (Dumuid et al., 2018), microbiome analysis (Lin et al., 2014), and nutritional epidemiology (Leite, 2016).

Examples of problems with both compositional outcomes and explanatory variables include relating the percentage of males and females with different education levels across countries (Filzmoser et al., 2018), modeling the relationship between age structure and consumption structure across economic areas (Chen et al., 2017), and understanding how different methods for estimating the composition of white blood cell types are related (Aitchison, 1986; Alenazi, 2019).

All current methods developed specifically for problems where both the outcome and the explanatory variable are compostional require data transformation. Chen et al. (2017) uses a log-ratio transformation for both the response and explanatory compositional variables, while Alenazi (2019) transforms just the compositional explanatory variable. Transformations, and specifically log-ratio based transformations, have long been used for compositional data, and have benefits such as allowing for tests for hypotheses of interest regarding compositional covariates (Aitchison and Bacon-Shone, 1984), and parametric modeling of convex combinations of compositions using logistic normal distribution (Aitchison and Bacon-Shone, 1999). However, transformation based models limit interpretability directly on the compositional space (Hron et al., 2012; Morais et al., 2018), especially when complex, but commonly used transformations such as the isometric log-ratio (ILR) transformation (Egozcue et al., 2003) are used. For the simpler additive log-ratio (ALR) transformation, regression parameters can be interpreted based on the *Aitchison geometry* (Aitchison, 1982, 1986, 1992) of perturbation ('addition') and power transform ('multiplication') of compositions. However, these interpretations rely on distributional assumptions (like logistic-normal) for the compositional outcome (Billheimer et al., 2001).

Furthermore, many transformations,including all of the log-ratio-based ones, do not allow for compositional data with 0s and 1s (Filzmoser et al., 2018). While there are transformation-free methods for analysis of compositional data with 0s and 1s, such as modeling the compositional data via a latent Gaussian model (Butler and Glasbey, 2008), these methods were not developed for the context of both the outcome and explanatory variables being compositional and rely on distributional assumptions.

In this manuscript, we postulate a simple estimating equation that directly relates the expected value of the compositional outcome as a linear function of the compositional explanatory variable. Our approach does not require any transformation of the data, or any distributional assumption beyond a first moment specification. The method naturally accommodates 0s and 1s, thus treating data on the interior of the simplex the same as data on the boundary. By linearly relating the outcome and explanatory variables, the parameters in our model are easily interpretable directly on the unit simplex in a distribution-free way, unlike transformation based compositional regression models. We develop an expectation-maximization (EM) (Dempster et al., 1977) algorithm for fast and accurate parameter estimation via constrained maximization of the quasi-likelihood that respects the unit sum nature of the compositional data. We present simulation results comparing the models for compositional data under a variety of data generating mechanisms. We also present permutation-based tests for assessing whether or not there exists a linear dependency between the outcome and explanatory variables, and for equality of effect sizes of two or more of the predictor components. We evaluate the operating characteristics of these tests via simulation.

Finally, we demonstrate the utility of our model with two data analyses from education and medical research.

2. Review of Transformation Based Compositional Regression Models

Current models for problems with compositional outcomes and explanatory variables rely on transforming the compositional data from $\mathbb{S}^D$ to $\mathbb{R}^{D-1}$. There are two recommended transformations for compositional data in the regression context. The first transformation is the additive log-ratio (ALR) transformation (Aitchison, 1986), where for $\mathbf{z} \in \mathbb{S}^D$,

$$alr(\mathbf{z})_j = ln\left(\frac{z_j}{z_D}\right),\ j = 1, \ldots, D-1.$$

The second transformation is the ILR transformation (Egozcue et al., 2003; Hron et al., 2012; Filzmoser et al., 2018), where

$$ilr(\mathbf{z})_j = \sqrt{\frac{D-j}{D-j+1}} ln\left(\frac{z_j}{\left(\prod_{k=j+1}^{D} z_k\right)^{\frac{1}{D-j}}}\right),\ j = 1, \ldots, D-1.$$

Both of these transformations allow the transformed compositions to be used as covariates and outcomes in a standard linear regression model without having to constrain the regression coefficients (Hron et al., 2012). While Van den Boogaart and Tolosana-Delgado (2013) note that the two transformations can be used interchangeably in terms of producing similar models, each one has its pros and cons.

The ILR transformation is isometric, in that it preserves angles and distances, but is more challenging to interpret (Egozcue et al., 2003; Filzmoser et al., 2018; Van den Boogaart and Tolosana-Delgado, 2013). The ILR model presented by Chen et al. (2017) assumes that for an outcome $\mathbf{y} \in \mathbb{S}^{D_r}$ and explanatory variable $\mathbf{x} \in \mathbb{S}^{D_s}$, that

$$E[ilr(\mathbf{y})_k|\mathbf{x}] = \beta_{0k} + \sum_{j=1}^{D_s-1} \beta_{jk} ilr(\mathbf{x})_j,\ k = 1, \ldots, D_r - 1. \tag{1}$$

Here β_{11} has an interpretation as the effect of increasing the relative value of x_1 by 1 compared to the rest of $\mathbf{x}$, holding the ratios between the other components of $\mathbf{x}$ constant, on the change of the relative value of y_1 compared to the rest of $\mathbf{y}$; the other regression coefficients have no meaningful interpretation (Hron et al., 2012; Chen et al., 2017). To obtain the effects of relative changes of each part of $\mathbf{x}$ on $\mathbf{y}$, one must use the permutation operation $\mathbf{z}^l = (z_l, z_1, \ldots, z_{l-1}, z_{l+1}, \ldots, z_D)$, and estimate $D_r \cdot D_s$ separate models where

$$E[ilr(\mathbf{y}^{l_1})_k] = \beta_{0k}^{(l_1,l_2)} + \textstyle\sum_{j=1}^{D_s-1} \beta_{jk}^{(l_1,l_2)} ilr(\mathbf{x}^{l_2})_j,\ k = 1, \ldots, D_r - 1,\ l_1 = 1, \ldots, D_r,\ l_2 = 1, \ldots, D_s. \quad (2)$$

The coefficients of interest would then be $\beta_{11}^{(l_1,l_2)}$ for each combination of l_1 and l_2 (Chen et al., 2017; Filzmoser et al., 2018). As parameter estimation is performed using standard maximum likelihood for linear regression models, this procedure is not computationally expensive. However, using multiple versions of a model to obtain a set of coefficients that cannot be interpreted jointly is undesirable.

There are two additional downsides. First, the ILR transformation does not allow for 0s in the compositions. Hence if either $\mathbf{x}$ or $\mathbf{y}$ are categorical or are compositions with 0's, this framework can not be used, even though categorical variables are still in the unit simplex. Second, the coefficients of interest can only be interpreted in terms of changes in the relative values of each part of the compositional data to the geometric mean. This does not permit for simple interpretation of the coefficients in terms of the direct effect of changing the value of $\mathbf{x}$ within the simplex on the expected value of $\mathbf{y}$ in the simplex (Morais et al., 2018). The lack of a simple interpretation for the coefficients in (2) have forced practitioners to instead rely on graphical techniques to display the estimated response surface of $\mathbf{y}$ as a function of $\mathbf{x}$ (Nguyen et al., 2018). One could attempt to remedy these issues by using the ALR transformation as it only needs to run one model and the regression coefficients are easier to interpret (Billheimer et al., 2001) in terms of perturbations of compositions. However from a practitioner's point-of-view, perturbations can be more challenging to interpret than direct change (differences) in response due to a change in the covariate, and can only be easily visualized for compositions with 3 or fewer components (Billheimer et al., 2001). Also the interpretations of ALR on the simplex rely on distributional assumptions (like logistic-normal) for the compositional outcome (Billheimer et al., 2001). Additionally, ALR does not preserve isometry, and does not allow for 0s and 1s. Finally, both log-transformations rely on the assumption of normality for the transformed compositional outcome, as well as linearity only in the transformed Euclidean space.

Alenazi (2019) takes a different approach to compositional regression, as only the explanatory compositional variable $\mathbf{x}$ is transformed. While Alenazi (2019) is more interested in prediction accuracy than interpretation and uses a complex principal components based transformation, one can use any transformation t (e.g., the ALR or ILR transformations). The assumed regression model is the pseudo-multinomial logit (pseudo-ML) specification (Papke and Wooldridge, 1996; Mullahy, 2015; Murteira and Ramalho, 2016):

$$\begin{aligned} E[y_k|\mathbf{x}] &= \frac{\exp(\beta_{0k} + \sum_{j=1}^{D_s-1} \beta_{jk} t(\mathbf{x})_j)}{1 + \sum_{k=1}^{D_r-1} \left[\exp(\beta_{0k} + \sum_{j=1}^{D_s-1} \beta_{jk} t(\mathbf{x})_j)\right]}, \; k = 1, \ldots, D_r - 1 \\ E[y_{D_r}|\mathbf{x}] &= \frac{1}{1 + \sum_{k=1}^{D_r-1} \left[\exp(\beta_{0k} + \sum_{j=1}^{D_s-1} \beta_{jk} t(\mathbf{x})_j)\right]}. \end{aligned} \tag{3}$$

Murteira and Ramalho (2016) discuss both quasi-maximum and maximum likelihood (QML and ML) methods for estimation of the coefficients. However, Alenazi (2019) uses a QML method which allows for 0 values in $\mathbf{y}$ (Papke and Wooldridge, 1996; Mullahy, 2015; Murteira and Ramalho, 2016), and does not make any distributional assumptions about $\mathbf{y}$.

Despite this psuedo-ML method allowing for potential 0s in $\mathbf{y}$ (and in $\mathbf{x}$ if one uses a transformation that allows for 0s, such as the α-transformation (Tsagris, 2015)), the regression coefficients are still only interpretable in terms of effects of changing a transformed version of $\mathbf{x}$ on $\log\left(\frac{E[y_j]}{E[y_{D_r}]}\right)$. In order to interpret the model in terms of changes within the simplex, one would again need to resort to graphical techniques.

3. Direct Linear Regression of Compositional Variables on the Simplex

For composition-on-composition regression that is interpretable without transformations and allows 0's in both $\mathbf{y}$ and $\mathbf{x}$, we directly model the expectation of $\mathbf{y}$ as a linear function of $\mathbf{x}$ as:

$$E[\mathbf{y}|\mathbf{x}] = \sum_{j=1}^{D_s} x_j \mathbf{b}_j, \tag{4}$$

where $\mathbf{b}_j$'s are D_y-dimensional vectors. Letting $\mathbf{B}$ represent the matrix with the jth row $\mathbf{B}_{j*} = \mathbf{b}_j^{'}$, we can rewrite the model in (4) as

$$E[\mathbf{y}|\mathbf{x}] = \mathbf{B}^{'}\mathbf{x} \ . \tag{5}$$

Because $\mathbf{y}$ is compositional, we require $\sum_{k=1}^{D_r} E[y_k|\mathbf{x}] = 1$ for all compositional $\mathbf{x}$. A necessary and sufficient condition for this is $\mathbf{B}$ to be a Markov (transition) matrix with non-negative entries and rows summing to 1, i.e.,

$$\mathbf{B} \in \{\mathbb{R}^{D_s \times D_r} | B_{jk} \geqslant 0, \sum_{k=1}^{D_r} \mathrm{B}_{jk} = 1 \text{ for } j = 1, \ldots, D_s\} \ .$$

This transformation-free model allows 0s and 1s in both $\mathbf{x}$ and $\mathbf{y}$ as (5). To see this, note that $\mathbf{B}'\mathbf{x}$ is well-defined for entire $\mathbf{x}$-simplex, including the boundary. Also, the model only makes a statement about the conditional expectation of $\mathbf{y}$ given $\mathbf{x}$, and not the distribution of the compositional outcome. Finally, the estimation procedure (defined later in (8)) is based on Kullback-Leibler loss using (5) which is well-defined even if the compositional responses has 0's.

We note that this model is formulated for compositional $\mathbf{y}$ and $\mathbf{x}$ that lie strictly on the unit simplex, i.e. $\sum_{k=1}^{D_r} y_k = 1$ and $\sum_{j=1}^{D_s} x_j = 1$. The term 'compositional data' is often used in a broader sense to simply represent a vector of positive numbers without the sum-to-one constraint (Van den Boogaart and Tolosana-Delgado, 2013). Modeling such general forms of compositions often relies on the principle of *scale invariance*, which states that "the information in a composition does not depend on the particular units in which the composition is expressed" (Filzmoser et al., 2018). To analyze such general compositional data with our method, one would have to normalize both the outcomes and predictors to lie on the unit simplex. Therefore, if applied to general compositional data, our model remains invariant to multiplication by a constant as we can always normalize the compositional data to lie in the unit simplex.

3.1 *Interpretability*

Our model allows us to directly estimate how the expected value of $\mathbf{y}$, rather than some transformed version of $\mathbf{y}$, is associated with changes in $\mathbf{x}$. This association between $\mathbf{x}$ and $E[\mathbf{y}]$ is expressed directly in terms of the regression coefficient matrix $\mathbf{B}$. If x_j increases by $\Delta \in (0, 1 - x_j]$, at the expense of x_k decreasing by Δ (assuming $x_k \geqslant \Delta$) and holding the rest of $\mathbf{x}$ constant, the expected change in $E[\mathbf{y}]$ is expressed as $\Delta(\mathbf{B}_{j*} - \mathbf{B}_{k*})$. This interpretation respects the fact that increasing one component of $\mathbf{x}$ necessarily involves the trade-off of decreasing at least one other component of $\mathbf{x}$. For example, if $\mathbf{x}$ represents the proportion of each day spent on different activities such as sleep, physical activity, and sedentary time, we may be interested in how components of a compositional $\mathbf{y}$ are expected to change when we increase physical activity and decrease sedentary time. We also may be interested in how this compares to the change of $\mathbf{y}$ when we instead increase physical activity at the expense of sleep (Dumuid et al., 2018). Another example application where this interpretation is useful is in marketing, where teams may want to know whether to increase the percentage of expenditure on television advertisements at the expense of radio advertisements or press advertisements in order to best increase their market share (Morais et al., 2018).

The simple interpretation of the direct regression model stands in stark contrast to the vague interpretation of the coefficients in the ILR model or any model which transforms $\mathbf{y}$ and/or $\mathbf{x}$. The interpretation of $\mathbf{B}$ is simple to communicate to non-statisticians without graphical techniques, does not require familiarity with the compositional transformations, and only requires estimating one single model for $E[\mathbf{y}|\mathbf{x}]$, rather than $D_r \times D_s$ models as for the ILR transformation. Even when compared to the ALR model, the interpretability of our model does not require any distributional assumption beyond a first moment assumption, whereas the ALR model interpretability relies on logistic-normal assumption. Also our model quantifies the regression coefficients directly in terms of change (difference) in the compositional outcome given change in the compositional covariate (respecting the sum-to-one constraint), which can be easier to interpret than the ALR regression coefficients explained in terms of perturbations on the simplex (Billheimer et al., 2001).

With our model, the hypothesis of linear independence, i.e. $E[\mathbf{y}|\mathbf{x}] = E[\mathbf{y}]$, is equivalent to restricting the model in (5) such that the rows of $\mathbf{B}$ are equal. We develop a row-permutation test for this hypothesis of interest in Web Appendix A, and show that it has well-calibrated Type-I error rates and high power.

As $\mathbf{B}$ is a Markov matrix, the rows of $\mathbf{B}$ are themselves members of $\mathbb{S}^{D_r}$. If we let $x_j = 1$, which means that $\mathbf{x}$ is in the jth corner of $\mathbb{S}^{D_s}$, (4) shows that $E[\mathbf{y}|x_j = 1] = \mathbf{b}_j = \mathbf{B}_{j*}$. For the case when $D_r = 3$, this means we can actually visualize the coefficients themselves using a ternary diagram (Hamilton and Ferry, 2018). Consider the following two values of $\mathbf{B}$:

$$\mathbf{B}^{(1)} = \begin{pmatrix} .90 & .05 & .05 \\ .05 & .90 & .05 \\ .05 & .05 & .90 \end{pmatrix} ; \ \mathbf{B}^{(2)} = \begin{pmatrix} .40 & .30 & .30 \\ .30 & .40 & .30 \\ .30 & .30 & .40 \end{pmatrix}$$

$\mathbf{B}^{(1)}$ represents the setting when $\mathbf{y}$ and $\mathbf{x}$ are highly correlated, while $\mathbf{B}^{(2)}$ represents the setting when $\mathbf{y}$ and $\mathbf{x}$ are weakly correlated. This interpretation is derived directly from the simple analytic interpretation of the direct regression model in (4). This is also seen through plotting the rows of these two matrices in a ternary diagram, as in Figure 1. Each number in the plot corresponds to a row in the two values of $\mathbf{B}$. The plot of $\mathbf{B}^{(1)}$ shows that $E[\mathbf{y}|\mathbf{x}]$ substantially changes with $\mathbf{x}$, as changes in $E[\mathbf{y}|\mathbf{x}]$ with $\mathbf{x}$ can be expressed as scaled differences in the rows of $\mathbf{B}$. However, the plot of $\mathbf{B}^{(2)}$ shows much smaller changes for

$E[\mathbf{y}|\mathbf{x}]$ with $\mathbf{x}$. Confidence regions for each row of $\mathbf{B}$ can also be plotted within the diagram. We demonstrate this in the example in Section 6.1.

[Figure 1 about here.]

3.2 *Amalgamation of categories*

In addition to the simple interpretation, the direct regression model in (4) exhibits other convenient statistical properties. First, consider the case when two rows, j_1 and j_2, of $\mathbf{B}$ are equal. This implies that increasing x_{j_1} at the expense of x_{j_2} does not change $E[\mathbf{y}|\mathbf{x}]$. We then have

$$E[\mathbf{y}|\mathbf{x}] = \sum_{j \neq j_1, j_2}^{D_s} x_j \mathbf{b}_j + x_{j_1}\mathbf{b}_{j_1} + x_{j_2}\mathbf{b}_{j_2} = \sum_{j \neq j_1, j_2}^{D_s} x_j \mathbf{b}_j + \mathbf{b}_{j_1}(x_{j_1} + x_{j_2}) \ , \tag{6}$$

which shows that we can treat the combined components $x_{j_1} + x_{j_2}$ as a single (amalgamated) category. This not only simplifies interpretation of the direct regression model, but also means that there is one less row of $\mathbf{B}$ to estimate. Because we do not know apriori whether any two rows of $\mathbf{B}$ are equal, we develop a column-permutation test for the null hypothesis of equality between any two rows of $\mathbf{B}$ in Web Appendix A. As with the test for linear independence, this test is also shown to have well calibrated Type-I error rates and has high power.

Similarly, the direct regression model can easily accommodate amalgamating components y_{k_1} and y_{k_2}. The direct regression model implies that

$$E[y_{k_1} + y_{k_2}|\mathbf{x}] = \sum_{j=1}^{D_s} B_{jk_1} x_j + \sum_{j=1}^{D_s} B_{jk_2} x_j = \sum_{j=1}^{D_s} (B_{jk_1} + B_{jk_2}) x_j. \tag{7}$$

Thus, conditional expectations of amalgamations of components of $\mathbf{y}$ can be obtained through amalgamating the corresponding columns of $\mathbf{B}$. This ensures that the model is invariant to amalgamation of outcome components. Rather than having to perform separate regressions for different choices of amalgamation of the outcome components, practitioners can simply perform one regression using the full set of components, and amalgamate columns of $\mathbf{B}$ post-hoc.

3.3 *Special cases*

We now demonstrate how our direct regression subsumes some commonly used models as special cases.

3.3.1 *Categorical covariates.* For each observation i, assume that the covariate of interest is whether or not the observation belongs to one of $j = 1, \ldots, D_s$ groups. If observation i belongs to subgroup j, we let $\mathbf{x}_i = \mathbf{e}_j$, where $\mathbf{e}_j$ is the compositional vector with a 1 in the jth index. We now have an ANOVA-like model, but with a compositional outcome.

This model has been considered in the literature where only the outcome is compositional, but previous solutions have either used an ILR transformation for $\mathbf{y}$ (Filzmoser et al., 2018) or assumed that $\mathbf{y}|\mathbf{x}$ follows a Dirichlet distribution (Maier, 2014). Our model allows for a transformation-free and distribution-free solution for this problem. The formulation of our model in (4) shows that $\mathbf{B}_{j*} = E[\mathbf{y}|\mathbf{x} = \mathbf{e}_j]$, i.e., the rows of $\mathbf{B}$ are simply interpreted as the expectation for the j^{th} group.

3.3.2 *Categorical outcome.* We now restrict $\mathbf{y}$ to be categorical, meaning that each observation i belongs to one of $k = 1, \ldots, D_r$ groups. The standard model for this case would be a multinomial logistic model, using the ALR or ILR transformed $\mathbf{x}$ as covariates (Filzmoser et al., 2018). However, we can use the model in (4), which allows for direct estimation of $E[y_k|\mathbf{x}] = P(\mathbf{y} = \mathbf{e}_k|\mathbf{x}),\ k = 1, \ldots, D_r$. This is equivalent to performing multinomial linear regression, with an identity link. The identity link is the canonical link here, as the covariates are compositional. Further restricting $\mathbf{x}$ to be categorical reduces this to a $D_r \times D_S$ contingency table. $\mathbf{B}_{j_1,k_1}$ can be interpreted now as the conditional probability $P(\mathbf{y} = \mathbf{e}_{k_1}|\mathbf{x} = \mathbf{e}_{j_1})$ and $\mathbf{B}_{j_1,k_1} - \mathbf{B}_{j_2,k_1}$ is the risk difference between groups.

This interpretation also gives insight into how one can decide on the number of compositional components to use. When both the covariates and outcome are categorical, B_{kj} would be estimated by

$$\widehat{B}_{kj} = \frac{\sum_i I(y_i = \mathbf{e}_k \ \& \ \mathbf{x}_i = \mathbf{e}_j)}{\sum_i I(\mathbf{x}_i = \mathbf{e}_j)}$$

. Letting $n_j = \sum_i I(\mathbf{x}_i = \mathbf{e}_j)$, the 95% confidence interval half-width for $\widehat{B}_{kj}$ using the normal approximation would be $1.96\sqrt{\frac{\widehat{B}_{kj}(1-\widehat{B}_{kj})}{n_j}}$. Thus, if one were to increase the number of components used for $\mathbf{x}$, then there must be cells with fewer observations, and thus a larger confidence interval for the corresponding entries in $\mathbf{B}$.

3.3.3 *Discrete time series transition probabilities.* A specific case of a categorical outcome and covariate is in estimating time-invariant transition probabilities for a first-order Markov process. An example of this class of problems is estimating the probability of firms or institutions transitioning between specific credit ratings (Jones, 2005). Observations may transition between $r = 1, \dots, R$ states. In the ideal case, for each observation unit i, we observe their discrete state $\mathbf{y}_{i,t}$ over times $t = 0, \dots, T$. We are then interested in estimating the probability that each observation moves to state j at time t, given that they are in state k at time $t-1$ (assuming transition probabilities are constant over time and between observation units). The interpretation of $\mathbf{B}$ from Sections 3.3.2 and 3.3.1 shows that if the covariate in (4), $\mathbf{y}_{i,t-1}$, and the outcome is, $\mathbf{y}_{i,t}$, then $\mathbf{B}_{jk} = P(\mathbf{y}_{i,t} = \mathbf{e}_j | \mathbf{y}_{i,t-1} = \mathbf{e}_k)$, which is exactly the transition probability we seek to estimate. The estimation procedure we outline in Section 4 will then coincide with the MLE of $\mathbf{B}$.

3.3.4 *AR(1) model for compositional data.* Rather than observing the states of each observation unit, we may only observe the percentage of observations in each state at each time. For example, Jones (2005) presents the case where for each year between 1984-2004, we only observe the percentage of commercial banks that belong to four different categories of credit quality. Our observed data is now the percentage of units in the different states at time t, $\mathbf{y}_t$. Specifically, y_{tj} is the percentage of observations belonging to state j at time t. Lee et al. (1970), MacRae (1977), and Jones (2005) have shown that $E[\mathbf{y}_t|\mathbf{y}_{t-1}] = \mathbf{B}'\mathbf{y}_{t-1}$, where $\mathbf{B}_{ij}$ is again defined as $P(\mathbf{y}_{i,t} = \mathbf{e}_j|\mathbf{y}_{i,t-1} = \mathbf{e}_k)$. Thus, the direct regression model in (5) can be used to estimate the individual transition probabilities, despite only observing aggregate data. For such settings, our model can be perceived as an AR(1) model for the compositional time series y_t.

We give additional examples in Web Appendix B of realistic data generating mechanisms where the compositional data arises as an aggregation of categorical data, and for which $E[\mathbf{y}|\mathbf{x}]$ is correctly specified by the direct regression model.

4. Parameter Estimation

4.1 *Estimating Equation Approach*

In order to estimate the entries of $\mathbf{B}$, we note that the model in (5) implies that $E[y_k|\mathbf{x}] = \sum_{j=1}^{D_s} B_{jk}x_j$. As we are only interested in the first moment of $\mathbf{y}|\mathbf{x}$, we use an estimating equation approach and seek a function $\ell(\mathbf{B}; \mathbf{y}, \mathbf{x})$ such that

$$E_{\mathbf{B}_0}\left(\frac{d\ell}{d\mathbf{B}}\right) = 0 \ ,$$

where $\mathbf{B}_0$ is the true value of $\mathbf{B}$. A function ℓ which achieves this, while also allowing for 0s in $\mathbf{y}_i$ and $\mathbf{x}_i$, is the Kullback-Leibler distance (KLD) between two compositional vectors — the observed $\mathbf{y}_i$ and $E[\mathbf{y}_i|\mathbf{x}_i]$ (Fiksel et al., 2020) — i.e.,

$$\begin{aligned} \ell &= \sum_{i=1}^{N} \text{KLD}(y_i \parallel E[y_i \mid x_i]) \\ &= -\sum_{i=1}^{N}\sum_{k=1}^{D_r} y_{ik} \log\left(\frac{E[y_{ik}|\mathbf{x}]}{y_{ik}}\right) \qquad (8) \\ &= -\sum_{i=1}^{N}\sum_{k=1}^{D_r} y_{ik} \log\left(\frac{\sum_{j=1}^{D_s} B_{jk}x_{ij}}{y_{ik}}\right). \end{aligned}$$

Letting $\mathcal{F} = \{\mathbf{B}; B_{jk} \geqslant 0, \sum_{k=1}^{D_r} \mathrm{B}_{jk} = 1\}$ be the parameter space for $\mathbf{B}$, minimizing (8) with respect to $\mathbf{B}$ is equivalent to maximizing the log-quasi-multinomial likelihood (Mullahy, 2015; Alenazi, 2019):

$$\begin{aligned}\min_{\mathbf{B}\in\mathcal{F}} \ell(\mathbf{B};\mathbf{x},\mathbf{y}) &= \min_{\mathbf{B}\in\mathcal{F}} - \sum_{i=1}^{N}\sum_{k=1}^{D_r} y_{ik}\log\left(\frac{\sum_{j=1}^{D_s} B_{jk}x_{ij}}{y_{ik}}\right) \\ &= \max_{\mathbf{B}\in\mathcal{F}} \sum_{i=1}^{N}\sum_{k=1}^{D_r} y_{ik}\log\left(\sum_{j=1}^{D_s} B_{jk}x_{ij}\right) \qquad (9)\end{aligned}$$

The multinomial quasi-likelihood belongs to the linear exponential family (Gourieroux et al., 1984) and minimizing (8) (or equivalently, maximizing (9)) produces a consistent estimator for $\mathbf{B}_0$ in (this has been showed for other parameteric forms of the regression function in Gourieroux et al., 1984; Papke and Wooldridge, 1996; Mullahy, 2015). When $\mathbf{y}$ is categorical (examples in Sections 3.3.2 and 3.3.3), the quasi-likelihood becomes the proper likelihood for multinomial distribution and the estimate of $\mathbf{B}$ becomes the MLE. More generally for compositional $\mathbf{y}$ and $\mathbf{x}$, Fiksel et al. (2020) show that (8) is convex with respect to $\mathbf{B}$, guaranteeing existence of a global minimum.

4.2 *An EM Algorithm for Maximizing the Objective Function*

Alenazi (2019) also uses an estimating equation approach via minimization of the KLD between the observed and expected values for the compositional outcome in (3). Because the form of the conditional expected value in (3) is that used in multinomial logistic regression, the coefficients are unconstrained and Alenazi (2019) utilizes the Newton-Raphson (Böhning, 1992) algorithm for maximizing the log-quasi-multinomial likelihood. However, in our model the parameter space for $\mathbf{B}$ is the space of all Markov matrices. Thus it is difficult to employ the Newton-Raphson algorithm to maximize (9) and a constrained optimization is required.

We instead develop an EM algorithm for parameter estimation by maximization of (9). We first present the algorithm for the special case where $\mathbf{y}_i$'s are categorical (Section 3.3.2). We introduce "missing" pseudo categories $\mathbf{x}_i^*$ such that $\mathbf{x}_i^*|\mathbf{x}_i \sim Multinomial(1, \mathbf{x}_i)$ and assume $\mathbf{y}_i|\mathbf{B}, x_{ij}^* = 1 \sim Multinomial(1, \mathbf{B}_{j*})$, thus using a proper likelihood for the outcome. We then arrive at the following likelihood of $\mathbf{y}|\mathbf{x}$ (marginalizing out the psuedo-categories $\mathbf{x}^*$):

$$\begin{aligned}
p(\mathbf{y}|\mathbf{B},\mathbf{x}) &= \prod_{i=1}^{N}\left(\sum_{j=1}^{D_s} p(x^*_{ij}=1)p(\mathbf{y}^*_i|\mathbf{B},x^*_{ij}=1)\right) \\
&= \prod_{i=1}^{N}\left(\sum_{j=1}^{D_s} x_{ij}\prod_{k=1}^{D_r}(B_{jk})^{y_{ik}}\right) \\
&= \prod_{i=1}^{N}\prod_{k=1}^{D_r}\left(\sum_{j=1}^{D_s} B_{jk}x_{ij}\right)^{y_{ik}} \qquad (10)
\end{aligned}$$

Taking the negative of log of (10) gives us the form of the objective function in (9). Letting $\mathrm{B}^{(t)}_{jk}$ denote the value of B_{jk} after iteration t, the expected complete log-likelihood becomes

$$Q(\mathbf{B}|\mathbf{B}^{(t)}) = \sum_{i=1}^{N}\sum_{j=1}^{D_2}\left[E[x^*_{ij}|x_{ij},y_{ik},\mathrm{B}^{(t)}_{jk}](log(x_{ij})+\sum_{k=1}^{D_1} y_{ik}log(\mathrm{B}_{jk}))\right].$$

Noting that the M-step will require finding

$$\max_{\mathbf{B}\in\mathcal{F}}\sum_{i=1}^{N}\sum_{k=1}^{D_r}\sum_{j=1}^{D_s} E[x^*_{ij}|x_{ij},y_{ik},\mathrm{B}^{(t)}_{jk}]y_{ik}\log(\mathrm{B}_{jk}) \ , \qquad (11)$$

we see that the terms in (11) for which $y_{ik}=0$ will not influence the maximization. Thus, rather than evaluating both $E[x^*_{ij}|x_{ij},y_{ik}=0,\mathrm{B}^{(t)}_{jk}]$ and $E[x^*_{ij}|x_{ij},y_{ik}=1,\mathrm{B}^{(t)}_{jk}]$, we only have to evaluate the latter term. We thus introduce weights $\pi^{(t+1)}_{ijk}$ for the E-step at iteration $t+1$ which are equal to $E[x^*_{ij}|x_{ij},y_{ik}=1,\mathrm{B}^{(t)}_{jk}]$:

$$\pi^{(t+1)}_{ijk} = \frac{x_{ij}\mathrm{B}^{(t)}_{jk}}{\sum_{j=1}^{D_s} x_{ij}\mathrm{B}^{(t)}_{jk}}, \ i=1,\ldots,N, j=1,\ldots,D_s, k=1,\ldots,D_r \ .$$

The expected complete log-likelihood is now

$$Q(\mathbf{B}|\mathbf{B}^{(t)}) = \sum_{i=1}^{N}\sum_{k=1}^{D_r}\sum_{j=1}^{D_s} y_{ik}\pi^{(t+1)}_{ijk}\log(\mathrm{B}_{jk}) \ ,$$

and the M-step from (11) becomes

$$\max_{\mathbf{B}\in\mathcal{F}} Q(\mathbf{B}|\mathbf{B}^{(t)}) = \max_{\mathbf{B}\in\mathcal{F}}\sum_{i=1}^{N}\sum_{k=1}^{D_r}\sum_{j=1}^{D_s} y_{ik}\pi^{(t+1)}_{ijk}\log(\mathrm{B}_{jk}) \ . \qquad (12)$$

Due to the fact that $\sum_{k=1}^{D_r}\mathrm{B}_{jk}=1$ for $j=1,\ldots,D_s$, we can recognize the constrained maximization in (12) equivalent to maximizing $j=1,\ldots,D_s$ weighted multinomial likelihoods.

This implies the following closed form M-step:

$$\mathrm{B}_{jk}^{(t+1)} = \frac{\sum_{i=1}^{N} y_{ik}\pi_{ijk}^{(t+1)}}{\sum_{k=1}^{D_r}\sum_{i=1}^{N} y_{ik}\pi_{ijk}^{(t+1)}},\ k = 1, \ldots, D_r,\ j = 1, \ldots, D_s\ .$$

Having developed an EM algorithm when we restrict the outcome $\mathbf{y}$ to be categorical, Theorem 1 now extends the EM algorithm to the general case when $\mathbf{y}$ is compositional:

Theorem 1: *Let $f(t) = \sum_{i=1}^{N}\sum_{k=1}^{D_r} y_{ik}\log\left(\sum_{j=1}^{D_s} B_{jk}^{(t)} x_{ij}\right)$ be the value of the objective function after iteration t of the EM algorithm with compositional outcomes $\mathbf{y}$, using the same E and M steps as when $\mathbf{y}$ is categorical. Then $f(t+1) - f(t) \geqslant 0$, with strict inequality if $Q(\mathbf{B}^{(t+1)}|\mathbf{B}^{(t)}) > Q(\mathbf{B}^{(t)}|\mathbf{B}^{(t)})$.*

A proof is provided in Web Appendix C. Theorem 1 allows use of the same EM algorithm for estimation of $\mathbf{B}$, despite the fact that our approach is likelihood-free and only specifies $E[\mathbf{y}|\mathbf{x}]$. As both the E-step and M-steps are available in closed form, the implementation of this EM-algorithm is extremely fast. The EM-algorithm can be further accelerated through use of the SQUAREM R-package (Du and Varadhan, 2020).

5. Simulation studies

We first perform simulations to compare the performance of the direct regression model with that of the log-ratio model of Chen et al. (2017) and the pseudo-ML model of Alenazi (2019) across situations when only one of the three models is correctly specified. To generate realistic data, we first fit each model to two datasets with a compositional outcome and explanatory variable: the Education dataset (Section 6.1) and the White Cells dataset (Section 6.2). These fitted coefficients are then used as the true coefficient values for each model when simulating data. Compositional covariates $\mathbf{x}_i$ ($i = 1, \ldots, N; N = 100, 250, 500, 1000$) were simulated independently such that $x_i \sim Dirichlet(1, 1, 1)$. For the log-ratio model, we simulate $ilr(\mathbf{y}_i)|\mathbf{x}_i \sim \mathcal{N}(E[ilr(\mathbf{y}_i)|\mathbf{x}_i], 1)$, and used $\mathbf{y}_i = ilr^{-1}(\mathbf{y}_i)$ as the compositional outcome. One could also use the ALR transformation instead of the ILR for model fitting, however, these usually produce statistically almost indistinguishable predictions, and hence only ILR was used as it was the transformation originally used by Chen et al. (2017). For the pseudo-ML model, we let $t(\mathbf{x}) = ilr(\mathbf{x})$. Because our direct regression model and the pseudo-ML model both directly specify $E[\mathbf{y}_i|\mathbf{x}_i]$, we used the coefficients for each model from the two datasets to obtain the true conditional expected values, and then simulated $\mathbf{y}_i|\mathbf{x}_i \sim Dirichlet(10 \cdot E[\mathbf{y}_i|\mathbf{x}_i])$ for each model.

Each of the three models were fit on the simulated data. To compare models, we generated a large, independent test set and obtained the true $E[\mathbf{y}_i|\mathbf{x}_i]$ for each observation. We then obtain the average KLD between the true and estimated conditional means in this independent set. This full process is repeated 10,000 times for every combination of N, true data generating mechanism, and dataset.

For ease of comparison, panel A in Figure 2 shows the log KLD for each simulation setting, averaged across all 10,000 simulations. Unsurprisingly, the correctly specified model performs the best in conditional mean estimation across almost all settings.

[Figure 2 about here.]

Interestingly, the log-ratio model appears to perform much worse when it is misspecified, as compared to the direct regression model and the pseudo-ML model. Overall, these results show that each of these models can be used to model compositional regression models, and that the KLD (either estimated on a test set or through cross-validation) is a valid metric for model comparison.

Because our direct regression model does not specify a likelihood for $\mathbf{y}|\mathbf{x}$, we also compare performance of the models across additional data generating mechanisms, where the first moment is correctly specified by the direct regression model. As in Section (5), we estimate the coefficients of the direct regression model on the same two datasets, and generate covariates $\mathbf{x}_i$ using a uniform Dirichlet distribution. We then generated $\mathbf{y}_i|\mathbf{x}_i$ using three data generating mechanisms, the first two of which were presented by Murteira and Ramalho (2016): the multinomial, and Dirichlet-multinomial (overdispersed Multinomial) distributions. Further details are given in Web Appendix D. The third data generating mechanism was the Logistic-Normal distribution, where $alr(\mathbf{y}_i)|\mathbf{x}_i \sim \mathcal{N}(alr(\mathbf{B}'\mathbf{x}_i), 1)$. Because $E[alr^{-1}(\mathbf{y})|\mathbf{x}] \neq \mathbf{B}'\mathbf{x}_i$, the direct regression model is only approximately correct in this scenario.

The fitted direct regression models are evaluated via KLD on a test set, as in the previous set of simulations. Panel B in Figure 2 shows that the direct regression model outperforms these models across the first two data generating mechanisms, when the direct regression model is exactly correct. When the outcome is generated using the Logistic-Normal distribution, and the direct regression model is only approximately correct, the models perform similarly to the log-ratio model while the pseudo-ML performs worse. In addition, while the KLD is similar for the direct regression method across all data generating mechanisms, the model performs slightly worse for the models with higher variance for the compositional outcome. We perform additional simulations in Web Appendix B that show the direct regression models again outperforms the other two models when the compositional data arises as an aggregation of categorical data. Overall, these results show the importance of correctly specifying the conditional mean for the compositional outcome, as each model is the most accurate only when correctly specifying the true data generation mechanism.

6. Applications

To show that our method can realistically use data to address scientific questions in an interpretable manner, we now apply our method to two datasets which have a compositional predictor and a compositional outcome.

6.1 *Educational status of mothers and fathers in European countries*

Parental educational attainment has a large effect on child outcomes (Dubow et al., 2009). Templ et al. (2011) provide a dataset in the `robCompositions` R package that contains the percent of fathers and mothers with low, medium, and high education levels in 31 European countries. The question of interest presented by Filzmoser et al. (2018) is how the percentage of fathers with a given education level relate to the percentage of mothers with different education levels, across the 31 countries. We let y_{ik} be the percentage of fathers with education level k (1 = low (pre-primary, primary or lower secondary education), 2 = medium (upper secondary education and post-secondary non-tertiary education), 3 = high (first stage of tertiary education and second stage of tertiary education)) (Eurostat, 2015) in country i, and x_{ij} be the percentage of mothers with education level j.

Fitting the model in (5) leads to the following estimate of $\mathbf{B}$:

$$\widehat{\mathbf{B}} = \begin{pmatrix} .91 & .05 & .04 \\ .00 & .91 & .09 \\ .00 & .14 & .86 \end{pmatrix}$$

which shows high correlation between the educational attainment status of fathers and mothers (independence test p-value=0). The coefficients and 95% confidence regions, obtained via bootstrap, are shown in Figure 3. There is noticeably more uncertainty in estimation of $\mathbf{B}_{3*}$ than in the other rows of $\mathbf{B}$. In addition, there is very little uncertainty in $\widehat{\mathbf{B}}_{2,1}$.

[Figure 3 about here.]

The analytical interpretation of $\mathbf{B}$ means that increasing the percentage of mothers with a medium level of education level by .10, while decreasing the percentage of mothers with a low level of education level by .10, is associated with a change in the percentage of fathers with low, medium, and high educational status of -.091, .086, and .005, respectively. Similar affects are seen for other changes of the percentage of mothers with a given educational status. Because the linear regression model is invariant to aggregating outcome categories, we can estimate that on average, a country with 100% of mothers with a low educational status would have 91% of fathers with a low educational status and 9% of fathers with a medium or high educational status.

To visualize the model fit, we first obtain predicted values for each of the father educational compositions, using leave-one-out cross-validation (LOOCV) (Friedman et al., 2001), based off the mother educational compositions in each country. Figure 4 shows the observed versus predicted percentage of fathers with each level of education, across the 31 countries. The predicted percentages are all very close to the observed percentages, showing that our simple model is not only interpretable, but also appears to fit the observed data well.

[Figure 4 about here.]

We also use the KLD between the observed $\mathbf{y}$ and predicted $\widehat{\mathbf{y}}$ compare our model to the log-ratio model and the pseudo-ML model, where $\mathbf{y}$ is estimated via LOOCV for all three methods. Each of the three methods had a KLD of .024, indicating similar model fit.

6.2 *White cell composition analysis*

Aitchison (2003) and Alenazi (2019) consider a dataset provided in the `ggtern` R package (Hamilton and Ferry, 2018) in which the proportions of white blood cell types (granulocytes, lymphocytes, and monocytes) in 30 blood samples are determined by both a time-consuming microscopic analysis and an automated image analysis. The microscopic analysis is known to produce accurate results, while the accuracy of the image analysis is unknown. If the estimated compositions from the microscopic analysis can be predicted by the compositions estimated by the image analysis, it would be time-saving to use the automated image analysis in the future.

We let y_{ik} and x_{ij} be the estimated composition of white blood cell type k and j (1 =

granulocytes, 2 = lymphocytes, 3 = monocytes) by the microscopic and image analysis, respectively. The estimate of **B** from our direct regression is

$$\widehat{\mathbf{B}} = \begin{pmatrix} .97 & .03 & .00 \\ .00 & 1.00 & .00 \\ .00 & .04 & .96 \end{pmatrix}$$

which shows extremely high correlation between the compositional outcome and explanatory variables (independence test p-value ≈ 0). An increase in the estimated percentage of lymphocytes by .10 from the image analysis, at the expense of a .10 decrease of the estimated percentage of monocytes, is associated with a change in the estimated proportions of granulocytes, lymphocytes, and monocytes of 0, .096, and -.096, respectively, from the microscopic analysis. Figure 5 again shows that our method produces extremely accurate predictions, obtained via LOOCV.

[Figure 5 about here.]

Finally, we again compare our method to the methods presented in Section (2) using the KLD. As in the analysis in (6.1), the models perform nearly identically, with direct regression model and the log-ratio model producing a KLD of .005, and the pseudo-ML model producing a KLD of .006. These two analyses show that our method is not only more interpretable, but also comes without loss of fidelity to the observed data.

7. Discussion

In this manuscript, we have introduced a simple and novel direct regression model for compositional outcomes and explanatory variables that is fundamentally different from the existing suite of transformation-based methods for such problems. This direct regression model offers a simple interpretation of the regression coefficients, as opposed to the transformation-based methods. This simple interpretation will facilitate the use of this model by practitioners who are not deeply familiar with complex compositional data transformations like the *ilr* transformation, without having to resort to graphical techniques for visualizing the response surface. In addition to its simplicity, the direct regression model accommodates 0s and 1s in the data, seamlessly allows amalgamation of categories for both the covariate and the outcome, and subsumes common structures such as 2-way contingency tables and discrete time first order Markov processes. The estimating equations approach makes the model robust to misspecified data distributions. Fast parameter estimation is obtained through a likelihood-free EM algorithm. Permutation tests for global independence of $\mathbf{y}$ and $\mathbf{x}$, and amalgamation of categories in $\mathbf{x}$ are proposed. Analysis of two datasets demonstrated how our model can accurately approximate observed scientific data generating mechanisms. Finally, we have implemented our method in the publicly available R package `codalm` (Fiksel and Datta, 2020).

7.1 *Limitations*

We note that while our model offers several advantages over the log-ratio and pseudo-ML approaches, such as being transformation- and distribution-free, and accommodating 0s and 1s, there are also some limitations compared to these models.

Unlike log-ratio models, our model does not explicitly encode the concept of subcompositional coherence, which posits that the analysis results for a subcomposition will remain invariant to whether one models full composition or just the subcomposition without knowledge of the other components. Implicit in this property, is the assumption that the full composition and the sub-composition can be analyzed using the same class of models. Thus, if there is strong prior knowledge about subcompositional coherence for a specific application, in which case one should use the log-ratio based models. However, we note that many types of compositional data do not have subcompositional coherence. For example, any compositional data arising as aggregates of categorical data, or compositional data with 0's will generally not have this property. Our model, however, does have coherence with respect to amalgamation of outcome categories, as illustrated in (7).

Our model is linear and linearity can sometimes be a restrictive assumption. As discussed in Section 3, the linear model (4) for compositional outcome $\mathbf{y}$ and compositional covariate $\mathbf{x}$ implies that $\mathbf{B}$ is a Markov matrix, i.e., the rows of $\mathbf{B}$ lie in the unit simplex. So any change in $E[y_k|\mathbf{x}]$ associated with a change of Δ in component x_j must be less than or equal to Δ. For example, our model would imply that if $E[\mathbf{y}|\mathbf{x}] = (.2, .8)$ when $\mathbf{x} = (.5, .5)$, $E[y_1|\mathbf{x}]$ could be no more than .3 and no less than .1 when increasing x_1 from .5 to .6. If it is believed for an application, that in some parts of the covariate simplex, small changes in a component of $\mathbf{x}$ are associated with large changes in a component of $\mathbf{y}$, then our direct linear model is not a good choice for such data. One should consider other alternatives like the log-ratio based models or pseudo-ML which is linear on the transoformed variables but induces a non-linear relationship on the simplex. However, we have also demonstrated that for several natural data generation processes including the examples of Section 3.3 and Web Appendix B, our linear model is correctly or approximately correctly specified.

We also note that the log-ratio model and the pseudo-ML model are more general approaches that have the ability to include multiple covariates of mixed variable types in the model, whereas our model is a specialized approach currently only for a single compositional covariate.

We present a full comparison of the properties of each model in Table 1.

[Table 1 about here.]

7.2 *Areas for Future Research*

One important future direction is developing a robust workflow for model comparison and selection for compositional regression problems. Although we have shown the potential of comparing the estimated KLD between models, there may be additional graphical and analytical tools that may yield better insight. Another important future direction is extending the direct regression model to allow for either continuous covariates or multiple compositional covariates, while maintaining simple interpretations for the compositional covariate coefficients. Log-ratio transformation based approaches for this problem simply extend the Chen et al. (2017) model by including the continuous covariates in the model (Morais et al., 2018). A potential solution is to use the direct regression model to model the partial dependence (Greenwell, 2017) between the compositional outcome and the compositional covariates of interest.

Data Availability Statement

The data that support the findings of this paper are openly available in a GitHub repository at `https://github.com/jfiksel/compregpaper`.

References

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

Aitchison, J. (1986). *The Statistical Analysis of Compositional Data.* Chapman & Hall,, London.

Aitchison, J. (1992). On criteria for measures of compositional difference. *Mathematical Geology* **24,** 365–379.

Aitchison, J. (2003). *The Statistical Analysis of Compositional Data.* Blackburn Press.

Aitchison, J. and Bacon-Shone, J. (1984). Log contrast models for experiments with mixtures. *Biometrika* **71,** 323–330.

Aitchison, J. and Bacon-Shone, J. (1999). Convex linear combinations of compositions. *Biometrika* **86,** 351–364.

Alenazi, A. (2019). Regression for compositional data with compositional data as predictor variables with or without zero values. *Journal of Data Science* **17,** 219–237.

Billheimer, D., Guttorp, P., and Fagan, W. F. (2001). Statistical interpretation of species composition. *Journal of the American statistical Association* **96,** 1205–1214.

Böhning, D. (1992). Multinomial logistic regression algorithm. *Annals of the institute of Statistical Mathematics* **44,** 197–200.

Butler, A. and Glasbey, C. (2008). A latent gaussian model for compositional data with zeros. *Journal of the Royal Statistical Society: Series C (Applied Statistics)* **57,** 505–520.

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

Dempster, A. P., Laird, N. M., and Rubin, D. B. (1977). Maximum likelihood from incomplete data via the em algorithm. *Journal of the Royal Statistical Society: Series B (Methodological)* **39,** 1–22.

Du, Y. and Varadhan, R. (2020). Squarem: An r package for off-the-shelf acceleration of em, mm and other em-like monotone algorithms. *Journal of Statistical Software, Articles* **92,** 1–41.

Dubow, E. F., Boxer, P., and Huesmann, L. R. (2009). Long-term effects of parents education on childrens educational and occupational success: Mediation by family interactions, child aggression, and teenage aspirations. *Merrill-Palmer quarterly (Wayne State University. Press)* **55,** 224.

Dumuid, D., Stanford, T. E., Martin-Fernández, J.-A., Pedišić, Ž., Maher, C. A., Lewis, L. K., Hron, K., Katzmarzyk, P. T., Chaput, J.-P., Fogelholm, M., et al. (2018). Compositional data analysis for physical activity, sedentary time and sleep research. *Statistical methods in medical research* **27,** 3726–3738.

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.

Eurostat (2015). Archive:living condition statistics - family situation of today's adults as children.

Fiksel, J. and Datta, A. (2020). *codalm: Transformation-Free Linear Regression for Compositional Outcomes and Predictors.*

Fiksel, J., Datta, A., Amouzou, A., and Zeger, S. (2020). Generalized Bayesian Quantification Learning. *arXiv e-prints* page arXiv:2001.05360.

Filzmoser, P., Hron, K., and Templ, M. (2018). *Applied Compositional Data Analysis With Worked Examples in R.* Springer, Cham, Switzerland.

Friedman, J., Hastie, T., and Tibshirani, R. (2001). *The elements of statistical learning,* volume 1. Springer series in statistics New York.

Gourieroux, C., Monfort, A., and Trognon, A. (1984). Pseudo maximum likelihood methods: Theory. *Econometrica: journal of the Econometric Society* pages 681–700.

Greenwell, B. M. (2017). pdp: an r package for constructing partial dependence plots. *The R Journal* **9,** 421–436.

Hamilton, N. E. and Ferry, M. (2018). ggtern: Ternary diagrams using ggplot2. *Journal of Statistical Software, Code Snippets* **87,** 1–17.

Hron, K., Filzmoser, P., and Thompson, K. (2012). Linear regression with compositional explanatory variables. *Journal of Applied Statistics* **39,** 1115–1128.

Jones, M. M. T. (2005). *Estimating Markov transition matrices using proportions data: an application to credit risk.* Number 5-219. International Monetary Fund.

Lee, T.-C., Judge, G. G., and Zellner, A. (1970). Estimating the parameters of the markov probability model from aggregate time series data.

Leite, M. L. C. (2016). Applying compositional data methodology to nutritional epidemiology. *Statistical methods in medical research* **25,** 3057–3065.

Lin, W., Shi, P., Feng, R., and Li, H. (2014). Variable selection in regression with compositional covariates. *Biometrika* **101,** 785–797.

MacRae, E. C. (1977). Estimation of time-varying markov processes with aggregate data. *Econometrica: journal of the Econometric Society* pages 183–198.

Maier, M. J. (2014). Dirichletreg: Dirichlet regression for compositional data in r.

Morais, J., Thomas-Agnan, C., and Simioni, M. (2018). Interpretation of explanatory variables impacts in compositional regression models. *Austrian Journal of Statistics* **47,** 1–25.

Mullahy, J. (2015). Multivariate fractional regression estimation of econometric share models. *Journal of Econometric Methods* **4,** 71–100.

Murteira, J. M. and Ramalho, J. J. (2016). Regression analysis of multivariate fractional data. *Econometric Reviews* **35,** 515–552.

Nguyen, T. H. A., Laurent, T., Thomas-Agnan, C., and Ruiz-Gazen, A. (2018). Analyzing the impacts of socio-economic factors on french departmental elections with coda methods.

Papke, L. E. and Wooldridge, J. M. (1996). Econometric methods for fractional response variables with an application to 401 (k) plan participation rates. *Journal of applied econometrics* **11,** 619–632.

Templ, M., Filzmoser, P., and Reimann, C. (2008). Cluster analysis applied to regional geochemical data: problems and possibilities. *Applied Geochemistry* **23,** 2198–2213.

Templ, M., Hron, K., and Filzmoser, P. (2011). *robCompositions: an R-package for robust statistical analysis of compositional data.* John Wiley and Sons.

Tsagris, M. (2015). Regression analysis with compositional data containing zero values. *arXiv preprint arXiv:1508.01913* .

Van den Boogaart, K. G. and Tolosana-Delgado, R. (2013). *Analyzing compositional data* *with R*, volume 122. Springer.

Supporting Information

Web Appendices A-D referenced in Sections 3, 4, and 5 are available with this paper at the Biometrics website on Wiley Online Library. The proposed methods are implemented in an R package `codalm` (https://cran.r-project.org/web/packages/codalm/index.html). All supporting code and data is available with this paper at the Biometrics website on Wiley Online Library.

Received October 2020

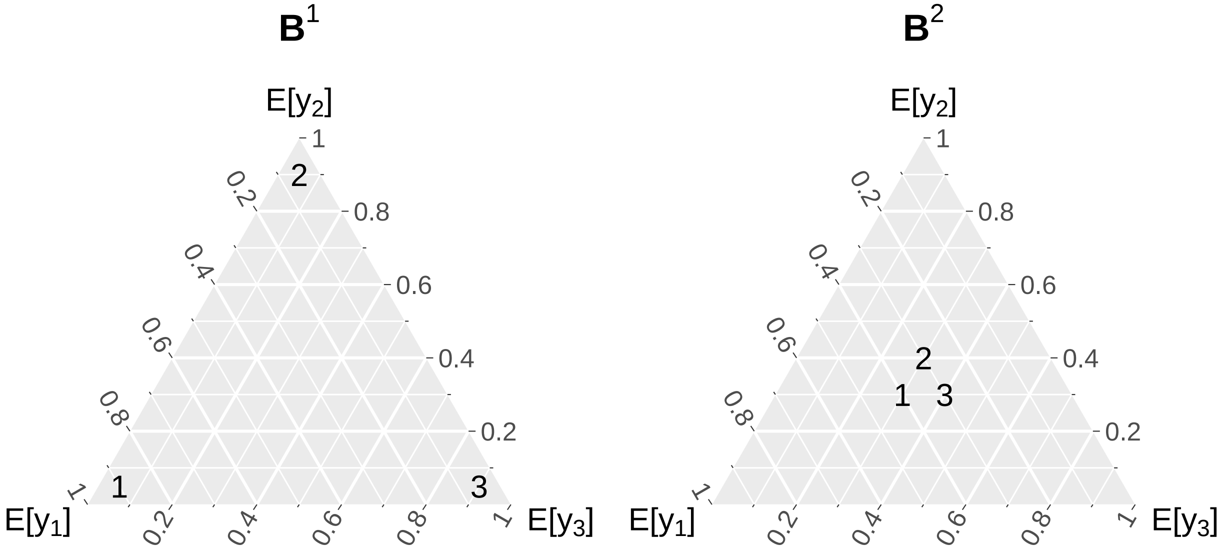

Figure 1: Visualization of the coefficients $\mathbf{B}$. For a number j, the point plots $\mathbf{B}_{j*}$ within a ternary diagram.

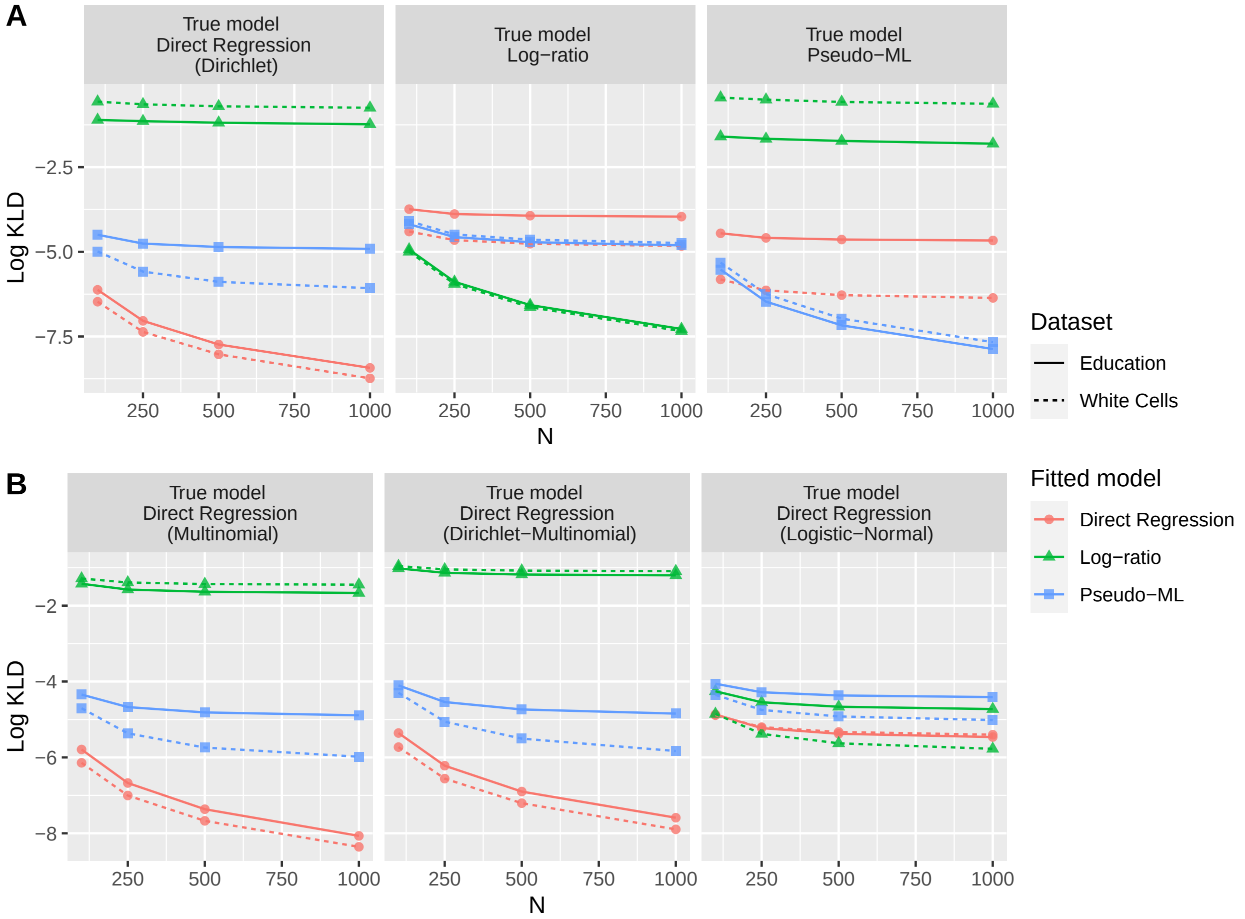

Figure 2: **Panel A:** Log KLD estimated using a test set, across various sample sizes and true models. Each column represents a different true model for the compositional outcome, with two sets of true coefficients values estimated on different datasets (solid and dashed lines). Each color and shape combination shows the estimated Log KLD based on the fitted model. **Panel B:** Log KLD estimated using a test set, when $E[\mathbf{y}|\mathbf{x}]$ is correctly specified by the direct regression model, across different data generating mechanisms. When the outcome data is generated by the Logistic-Normal distribution, the direct regression model is only approximately correct, thus leading to similar performance for all three models. This figure appears in color in the electronic version of this article, and any mention of color refers to that version.

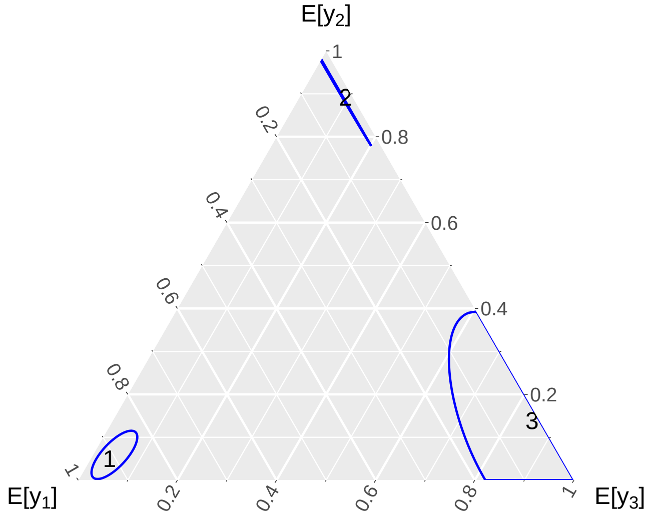

Figure 3: Visualization of the coefficients for regression the percentage of fathers of a given education level on the percentage of mothers of a given education level. Each row of $\widehat{\mathbf{B}}$ is labeled with a number in the ternary diagram. The 95% confidence region for each row is drawn in blue. This figure appears in color in the electronic version of this article, and any mention of color refers to that version.

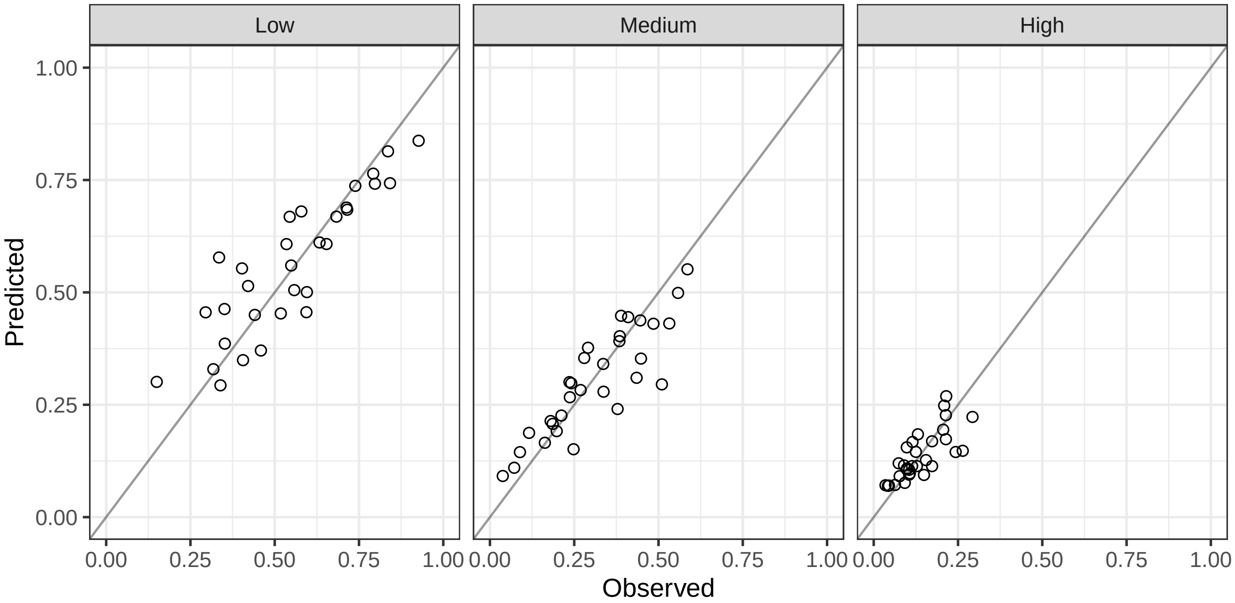

Figure 4: Observed versus predicted father educational attainment compositions across each of the 31 countries based on leave-one-out analysis. The grey line represents the identity line.

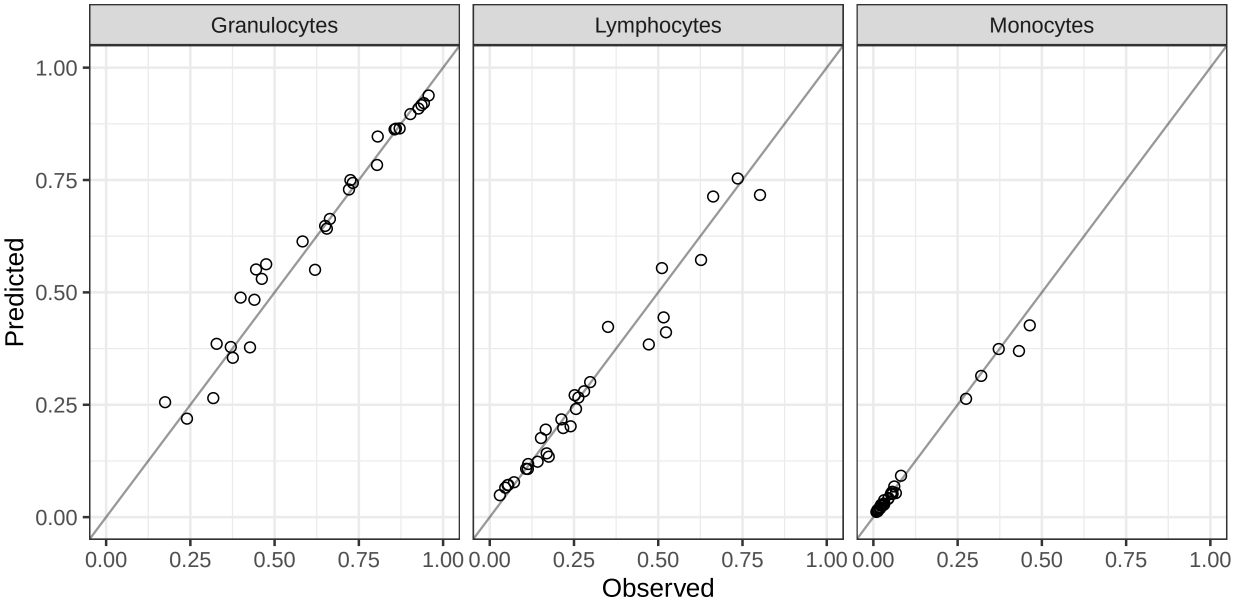

Figure 5: Observed versus predicted white blood cell composition estimates using the microscopic analysis from each of the 30 samples using leave-one-out analysis. The grey line represents the identity line.

Table 1: Comparison of properties between different models for compositional regression. A ✓ indicates that a model has the given property, while a ✗ indicates that a model does not have the given property.

Properties	Direct Regression	Log-ratio (ILR)	Log-ratio (ALR)	Pseudo-ML
Transformation-free	✓	✗	✗	✗
Accommodates 0s and 1s in both outcome and predictor compositions	✓	✗	✗	✓
Coefficients interpreted in terms of changes of $\mathbf{y}$ in the simplex	✓	✗	✗*	✗
Coefficients interpreted in terms of changes of log ratios of $\mathbf{y}$	✗	✓	✓	✓
Only requires running 1 model, instead of $D_r \times D_s$ models	✓	✗	✓	✗
Can be extended to include multiple covariates that may be compositional, continuous, or discrete	✗	✓	✓	✓
Explicitly encodes subcompositional coherence	✗	✓	✓	✗
Does not rely on distributional assumptions	✓	✗	✗	✓

*However, the regression coefficients for the ALR model can be interpreted in terms of pertubations of compositions in the simplex