Adapting Multiple Imputation for Compositional Survey Data

Sana Gupta, Benjamin Stockton, & Ofer Harel

Department of Statistics, University of Connecticut, Storrs, CT

https://doi.org/10.33697/ajur.2025.146

*Students: sana.gupta@uconn.edu**
Mentors: benjamin.stockton@uconn.edu, ofer.harel@uconn.edu

ABSTRACT
Compositional data, where each component is a proportion of a whole, presents unique statistical challenges, particularly when incomplete. Multiple Imputation (MI) has become a standard method for imputing incomplete quantitative, ordinal, or categorical data, but there are not any proposed imputation methods for incomplete compositional data that are able to preserve the characteristics of the compositions. We propose methods for imputing compositional data, and use the imputed datasets to conduct analysis on exercise motivation survey data. The novel method will be used to impute missingness in the original dataset, which serves as the basis for the model development. The results of the analysis will be used to evaluate the performance of our proposal against standard methods.

KEYWORDS
Applied Bayesian Statistics; Exercise Motivation; Missing Data; Multiple Imputation; Multivariate Statistics; Survey Methodology

INTRODUCTION

Missing data arises from various sources such as incomplete responses to surveys and errors during data collection. In surveys or measurements where data is compositional, the values for each compositional variable need to sum to a whole, typically 1 or 100%. **Figure 1** is a visual representation of how missingness can look in a dataset that is made up of compositional data. Each bar represents one participant or observation, and each colored section is a different compositional variable. Compositions need to add up to the same sum for each observation, and the lengths of the bars in the figure are constant to represent this relationship. The missingness, identified in the figure by the white gaps outlined by dotted lines, can occur differently from observation to observation. Some observations might have only one missing value like Participant 3, while others have all but one value missing, like Participant 2. Regardless of the amount of missingness present, the sum of all the values in the column (both known and unknown) is held at the known value for the whole. It is important to resolve missingness in order to proceed with most types of analysis.

For any missing value, there is inherent uncertainty about what each of the missing values could be. As a result, it is important that any method used to handle this missing data is able to consider the uncertainty when imputing, or filling in, the missing values. One commonly referenced method for handling missing data is Multiple Imputation (MI), which involves performing several imputations on each missing value to account for the variability that is present.[1, 2] However, conventional MI implementations does not take into account the need to maintain a specific sum across the whole observation.

Within MI, there are several widely-accepted imputation methods for standard numerical data, such as imputation by Bayesian linear regression,[3] predictive mean matching,[4] and random forests.[5] However, there is not currently a pro-

posal for imputing incompletely observed compositions in the statistical literature. Compositional data is a type of multivariate data in which each variable, or component, is part of a whole.[6] Each component individually represents a proportion, and the values in each column are relative to each other. Rather than the typical Euclidean space R^D, the sample space for compositional data is the simplex, a $D-1$ dimensional subspace of R^D due to the summation constraint.[7] These factors complicate the imputation of compositional data.

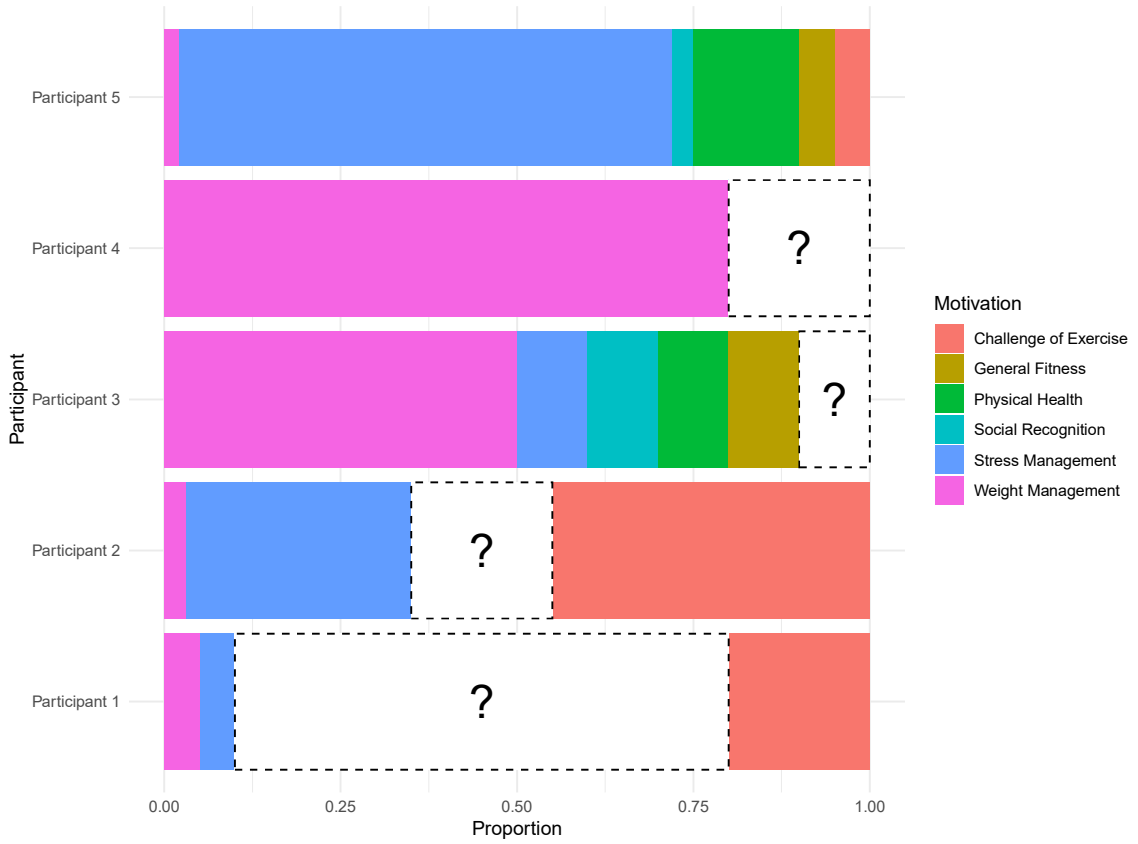

Figure 1. Each bar in this graph represents a participant, and the heights of each bar are the same to indicate a consistent total that needs to be reached. Missing values within a participant need to be imputed, which can be difficult when more than one component is missing.

In this paper, we propose a Scaling method and an Isometric Log-Ratio (ILR) Regression method for imputing incomplete compositional data. The Scaling method is based on the imputation of missing values at the composition component level using multiple imputation, and the subsequent rescaling of the data to maintain a sum of 1 for the whole observation. The ILR Regression method is based on the ILR transformation, which involves using the Isometric Log-Ratio transformation to transform the different components before imputing the missingness at the ILR level for each incomplete observation. With these methods, we address the imputation of missing compositional data. We are not aware of literature addressing the imputation of this particular type of data and we aim to fill that gap with the proposal of these methods.

These implementations will be evaluated in the context of a survey about exercise motivation, in which respondents have assigned percentages to the following exercise motivations in order to capture their relative importance:

1. Managing stress
2. Maintaining or improving weight and/or appearance
3. Gaining social recognition

4. Enjoying the challenge of exercising
5. Improving physical health
6. Maintaining or improving fitness

This paper is structured as follows. In the Background section, we provide information on the missing data analysis and compositional data analysis. In the Data section, we introduce the motivating dataset and discuss the data collection process. In the Methods section, we detail our proposed methods for imputation. In the Simulation section, we explain the process of our simulation study, and present the simulation results in the following section. In the Survey Data Analysis Results section, we present the results of our application of the methods to the motivating dataset. In the Discussion, we discuss the results and suggest potential future work to improve these methods.

BACKGROUND

Missing Data and Multiple Imputation

The preferred method for imputation can vary depending on the characteristics of the data generating process. These characteristics include the missingness mechanism and the nonresponse pattern, both of which concern the underlying reasons for and the distribution of missing data within the dataset. Rubin established a framework for treating missingness as probabilistic.[3] The missingness is represented by an indicator variable matrix R, where $R = 1$ when the value is complete and $R = 0$ when the value is incomplete.[8] The complete dataset, Y_{com}, is made up of its observed (Y_{obs}) and missing (Y_{mis}) parts such that $Y_{com} = (Y_{obs}, Y_{mis})$.[9] **Figure 2** illustrates this process, where each row is an observation and each column is a composition that sums to 1. In this process, any observation with more than two missing values requires multiple imputations because there are multiple possibilities for what each of those values could be. However, the observation with only one missing value does not require multiple imputations because there is only one possible value that allows the whole observation to sum to 1. One of the three following missingness mechanisms can be used to describe the pattern and cause of missingness in a dataset[10]:

Data are missing completely at random (MCAR) when the probability of being missing is not related to the data:

$$MCAR : P(R|Y_{\text{obs}}, Y_{\text{mis}}) = P(R).$$

Data are missing at random (MAR) when the probability of missingness is related to observed data, but is not related to unobserved data:

$$MAR : P(R|Y_{\text{obs}}, Y_{\text{mis}}) = P(R|Y_{\text{obs}}).$$

Data are missing not at random (MNAR) when neither of the above cases are true:

$$MNAR : P(R|Y_{\text{obs}}, Y_{\text{mis}}) \neq P(R|Y_{\text{obs}}).$$

Currently, Multiple Imputation (MI) is a widely used and theoretically grounded method for handling incomplete data.[11] Developed by Donald B. Rubin in the 1970s, MI addresses the uncertainty present in missingness by completing the dataset multiple times, resulting in several completed datasets that can be analyzed using standard analysis methods.[3] Unlike methods that only consider complete observations, Multiple Imputation uses information from both incomplete and complete observations, leading to greater efficiency.[12]

The multiple imputation framework involves three steps. First, a set of values is imputed for each of the missing data points, resulting in multiple complete datasets. This allows uncertainty about the missing data to be considered. After imputing, the analysis of interest is conducted on each of the complete datasets. The results from each completed dataset are then pooled using Rubin's rules, leading to a single estimate.[3]

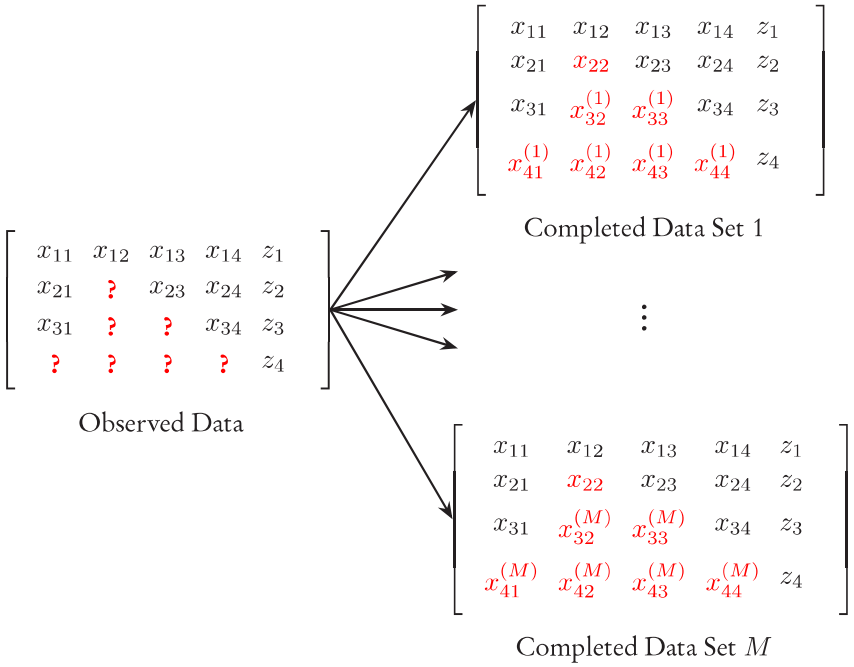

Figure 2. This diagram illustrates how the observed data (on the left) are imputed M times to create M completed data sets (on the right). The data x_{ik} for $i = 1, \ldots, 4$ and $k = 1, \ldots, 4$ are compositional data with four components observed on four study units. The data z_i are some additional completely observed variable hypothesized to be related to the compositions. Missing observations are highlighted in red text and labeled as "?" in observed data and replaced with imputations on the right as $x_{ij}^{(m)}$. Note that x_{22} does not have a completed data superscript since it is imputed only once. The value of x_{22} is known even if initially unobserved due to the geometric constraints of the sample space, i.e. the simplex, such that $x_{22} = 1 - \sum_{k \in \{1,3,4\}} x_{2k}$.

There is an abundance of literature on the imputation of conventional numerical data,[1, 9] and there are several software tools in programming languages such as R designed to work with incomplete data.[13] For example, the R package MICE (Multivariate Imputation by Chained Equations) is a popular choice for multivariate missing data.[14] The package sequentially imputes the incomplete variables to complete the dataset multiple times, in order to efficiently sample from the marginal posteriors of the missing data given the observed data, using cycles of univariate imputations for each incomplete variable.

Compositional Data

Compositional data is a subtype of multivariate numerical data where each value in the observation is a proportion or percentage, and the values must add up to a whole. While this particular behavior is similar to that of the multinomial distribution, compositional data is often made up of continuous proportions or percentages that may be dependent from observation to observation, rather than multinomial data where one must assume independent trials with a fixed number of total observations. Unlike standard numerical data, the summation constraint complicates the algebraic manipulation of compositional data. For example, if one value decreases, others must increase to maintain the sum. **Figure 3** shows responses of the first 15 subjects with complete responses in our dataset.

Compositional data is often discussed in the context of the D-part Aitchison simplex, a geometric framework for maintaining compositional data characteristics.[15] The unit simplex is the sample space for the compositional data. When all but one composition value is fixed, the final component is determined to maintain the total sum of one:

$$\text{Simplex}(S^{D-1}) = \left\{ (x_1, x_2, \ldots, x_D) \in R^D \mid x_i \geq 0, \sum_{i=1}^{D} x_i = 1 \right\}, \qquad \textbf{Equation 1.}$$

where S^{D-1} represents the $D-1$ dimensional simplex embedded in the R^D space.[6]

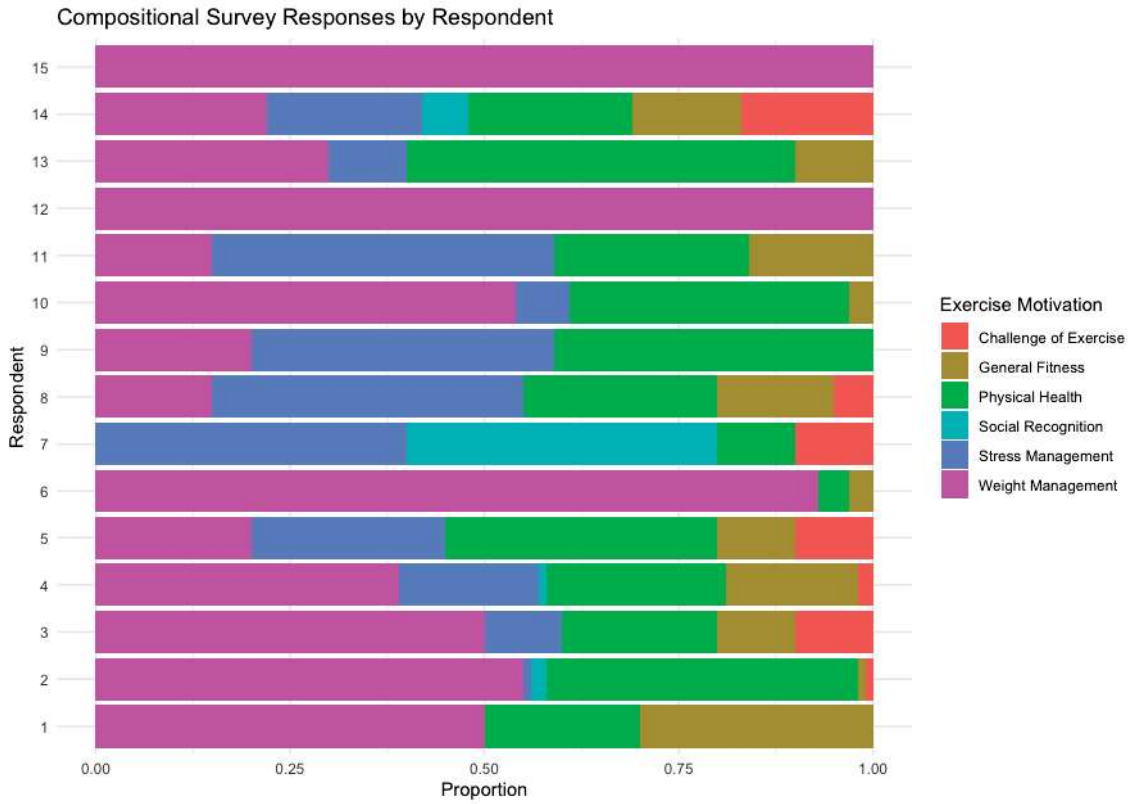

Figure 3. This bar chart shows the compositional responses to the exercise motivations survey for the first 15 respondents.

Log-Ratio transformations are a tool in compositional data analysis because they make the characteristics of compositional data easier to work with mathematically. One such characteristic is the dependency across each of the compositional parts. This dependency can result in collinearity when the compositional values are used as predictors in a model. For example, the Centered Log-Ratio (CLR) transformation is a common transformation used for compositional data because it maps original columns directly to transformed columns, making analysis and interpretation straightforward. The CLR transformation also results in a vector for each observation that sums to zero.[16] This causes collinearity to arise when fitting a linear model using all of the compositions. For this reason, we chose to work with the Isometric Log-Ratio transformation, defined as:

$$\mathrm{ILR}(x) := V^T \mathrm{clr}(x), \qquad \textbf{Equation 2.}$$

where V represents the matrix with columns that make up an orthonormal basis of the CLR-plane, V^T represents the transpose of matrix V, and $\mathrm{clr}(x)$ represents the Centered Log-Ratio transformation on a composition x defined in Equation 3.[16]

$$\mathrm{clr}(x) := \left(\log\left(\frac{x_1}{g(\mathbf{x})}\right), \log\left(\frac{x_2}{g(\mathbf{x})}\right), \ldots, \log\left(\frac{x_D}{g(\mathbf{x})}\right)\right)^T, \qquad \textbf{Equation 3.}$$

where $g(\mathbf{x})$ is the geometric mean of the components of $\mathbf{x}$,

The ILR transformation results in a vector that does not sum to a constant, allowing for a model to be computed without the issue of collinearity. The ILR data transformation allows D-part compositions to be represented as vectors unconstrained in R^{D-1}, resulting in data that are easy to analyze with most methods. The ILR transformation is useful because it allows us to circumvent issues of collinearity when building an analysis model.[16]

DATA

The motivating dataset comes from a survey of 340 participants about their motivations for exercising and their associated characteristics and demographic information. Age was the only demographic variable measured as an ordinal value. The participants' ethnicity, race, gender, income, and education levels were measured as categories or grouped into bins, with distributions as described in **Table 1**.

Factor	Group	Frequency
Age	Median	32
	Range	56 (18 – 74)
	IQR	17 (24 – 41)
Income	< $20,000	46
	$20,000 – $29,000	31
	$30,000 – $39,000	30
	$40,000 – $49,000	40
	$50,000 – $59,000	32
	$60,000 – $69,000	36
	$70,000 – $79,000	23
	$80,000 – $89,000	14
	$90,000 – $99,000	16
	> $100,000	62
Ethnicity	Not Hispanic/Latino	305
	Hispanic/Latino	23
	Prefer not to answer	10
Race	Black or African-American	48
	White	232
	Native American or Alaska Native	3
	Native Hawaiian or other Pacific Islander	27
	Asian	20
	More than one race	8
Gender Identity	Cisgender	320
	Transgender/Non-Binary	18
Gender	Male	112
	Female	210
	Non-Binary	16
Education	Less than high school	2
	High school diploma/GED	48
	Some college or technical/vocational school	114
	College graduate	98
	Some graduate school	10
	Graduate degree	66

Table 1. Frequencies of participant demographics.

The exercise motivations were measured using two different survey methods:

Survey Method 1 (EMI-2): The second volume of the Exercise Motivations Inventory (EMI-2) asks respondents to rate their agreement with questions about specific exercise motivations on a scale from 1 ("not true at all for me") to 5 ("very true for me"). These result in ordinal values, which are simpler to process, impute, and analyze. The developers of the EMI-2 questionnaire assigned the 51 questions to 14 distinct categories to summarize the responses.[17] Previous

work further narrowed these categories into six categories that broadly capture the motivations people have for exercising.

Survey Method 2 (Pie Chart Items): Survey participants were asked to indicate what percentage of their total exercise motivation can be attributed to each of six categories. The responses are required to sum to 100%, representing the individual's complete set of reasons for exercising. This method results in compositional data where the value of a response for one category must be considered relative to the others.

METHODS AND PROCEDURES

The first proposed method is the Scaling method. This method employs multiple imputation using the Predictive Mean Matching (PMM) technique to estimate the missing values at the composition component level. The PMM method involves gathering a list of candidate imputations from the closest observed responses to the predicted value for the missing response based on a Bayesian multiple linear regression model. One observation is randomly selected from the set of candidates, and the missing value is filled in using that observation.[18] During each iteration of the imputation step, we first impute on the compositional proportions, and normalize the values across each row by dividing them by the sum of the values in the row. This process ensures that the scaled values maintain relative proportions within each individual's compositional responses and sum to 1, preserving the structure of the data. Then, we perform the Isometric Log-Ratio (ILR) transformation on the completed compositions so we can use them in the analysis model.[16] This process is summarized in the top part of **Figure 4**.

The second novel method is the Isometric Log-Ratio Regression method. First, we perform an Isometric Log-Ratio transformation (ILR) on the compositional columns. The ILR transformation requires a fully observed compositional observation. To achieve this, the missing values are initially imputed with the mean value of the column. True zero values in the dataset are also replaced with a small constant to avoid undefined logarithms. The use of these mean values and small constants does not impact the forthcoming computation because these placeholder values do not remain in the data after this pre-processing step is complete. The `mice` function from the MICE R package then iterates through ten cycles of imputations to converge to the predictive distribution for the missing data. The ILR transform maps compositions from the D-part Aitchison-simplex to a $D-1$ dimensional Euclidean vector, as noted in Equation 2.[16]

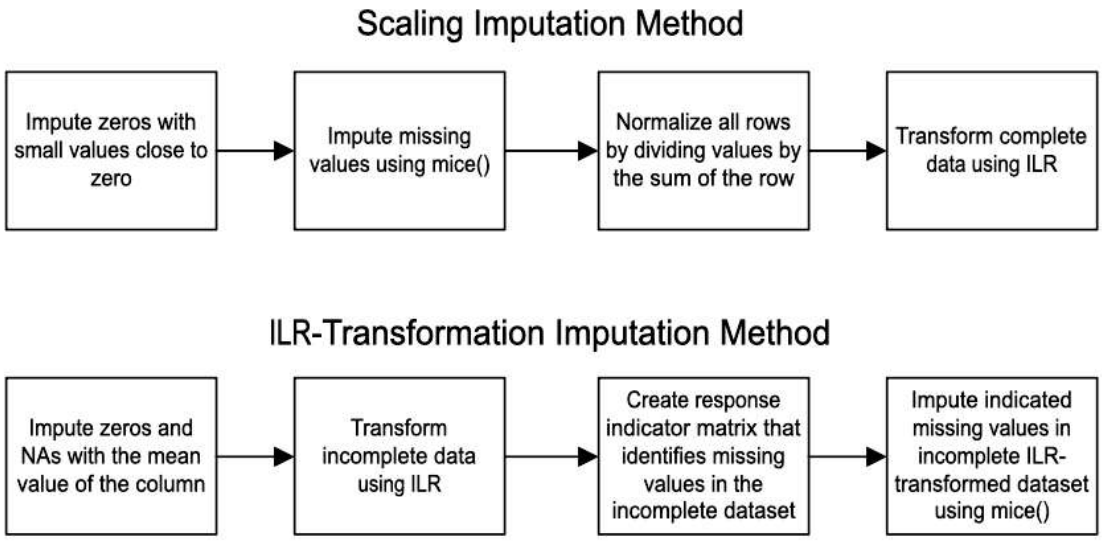

Figure 4. Flowchart detailing the steps for the Scaling and ILR-transformation imputation methods.

Then, we combine a binary response matrix to indicate the locations of missing values in the original incomplete dataset, and the demographic information with the pre-procesed ILR-transformed data. Multiple imputations are then performed

on the combined dataset using imputation by Bayesian multiple linear regression. The ILR transformed data are multivariate normal.[19] Thus, performing MICE using Bayesian linear regression allows the imputation distribution to converge to a valid joint multivariate normal distribution.[20] Finally, the function returns the imputed datasets. This process is summarized in the bottom part of **Figure 4**.

The ILR Regression method has theoretical advantages over the Scaling method, as it is based on the Isometric Log-Ratio transformation, which allows for appropriately handling compositional data by preserving the relative structure between each of the components. On the other hand, the Scaling method isn't based on any particular theoretical framework, making it more of a heuristic approach. The Scaling method may perform well empirically, as it is forcing each observation to sum to one in an intuitive way.

Simulation

We considered three methods for incomplete data analysis in comparison to analysis with the complete dataset. Each method results in a dataset that is complete and usable for analysis. The complete dataset created in the data generation step was used as the benchmark for the other methods.

For the first method, a dataset is created using the Complete Case Analysis (CCA) method[18]. When applied to a data set with compositional data, each row i with an incomplete composition, *e.g.* $(x_{i1}, ?, ?, ?, x_{i5}, x_{i6})'$ is deleted. The row-wise deletion removes the entire study unit or participant from the data set. While CCA is a simple and easy to implement ad hoc missing data analysis technique, it is often not theoretically justifiable nor an efficient use of the available data so we would not recommend its use in general. However, CCA is used in the simulations as a way to provide a minimally acceptable lower bound on the performance of other methods in the simulations.

In order to evaluate the performance of the various methods, a simulation study was performed. The simulation study and analyses for this project were all conducted using the R programming language.[13] We conducted a Monte Carlo simulation where a new dataset was generated for each iteration, and the methods under consideration were applied. This type of simulation provides an empirical evaluation of the methods' inferential performance under the considered data-generating processes.[21,22] The steps of the simulation are as follows:

1. The simulated demographic variables are generated by sampling the empirical distributions for these variables in the original dataset.
2. The simulated compositional variables are generated by predicting ILR values using a linear model that was previously fit to the original data. Simulated data are drawn from the Aitchison distribution to generate a complete dataset with compositional and demographic values matching the structure of the original dataset.[16] The Aitchison Distribution generalizes the Dirichlet distribution and the additive log-normal distribution in order to model compositional data.
3. The simulated EMI variables are generated by predicting values using a linear model that incorporates the demographic and ILR-transformed composition data as predictors. Simulated data are drawn from the assumed univariate normal distributions from the fitted linear models.
4. Missingness is added to the simulated datasets. Four different missingness proportions ($p_{miss} = 0.1, 0.25, 0.5, 0.75$) using a MAR missingness mechanisms were applied for a combination of twelve different missingness patterns. Other MCAR and MNAR mechanisms were also evaluated.
5. The Complete Case Analysis, ILR Regression, and Scaling methods are applied to the dataset. For ILR Regression and Scaling, $M = 50$ imputations were done for each missing value.

6. For each method's completed data set, the following linear regression model is fit:

$$\begin{aligned}\text{stress}_i = \beta_0 + \sum_{k=1}^{5} \beta_k \times \text{ILR}_i + \beta_6 \times \text{age}_i \\ + \beta_7 \times \text{education}_i + \beta_8 \times \text{income}_i \\ + \beta_9 \times \text{race}_i + \beta_{10} \times \text{ethnicity}_i \\ + \beta_{11} \times \text{gender}_i + \beta_{12} \times \text{cisgender}_i + \epsilon_i.\end{aligned} \quad \textbf{Equation 4.}$$

7. The regression results are pooled by combining estimates and standard errors for β_k where $k = 0, \ldots, 12$ from the multiple imputed datasets according to Rubin's rules,[3] and the pooled summary is recorded for each of the four methods applied in the iteration.

The above steps illustrate the simulation process for one iteration. For each combination of missingness proportion and mechanism, a total of $N_{sim} = 1000$ iterations were performed.

Since MAR is the most common missing data assumption, we concentrate on this assumption. Results for MCAR and MNAR are in the appendix and discussed further in the Discussion section. In order to introduce potential bias in the CCA estimates, we use a MAR mechanism, a logistic model that induces missingness on ILR_2 that is associated with the right tail of the response variable Y, on the ILR_2 coefficient, which is defined by the equation:

$$\text{logit}(P(R = 0)) = \alpha_0 + 3Y$$

where Y represents the stress response variable, $\alpha_0 = -3\bar{Y} - \log(1/p_{miss} - 1)$ is based on the desired missingness proportion, and $\bar{Y}$ is the sample mean of the stress response. This MAR mechanism satisfies the necessary condition that the probability of missingness for the to-be amputed predictor ILR_2 has an association with response Y.[18, 22]

In the outcome regression model, the stress EMI variable is regressed on the compositions in ILR form and demographics. The stress EMI variable is representative of the relationships between the composition and the other EMI categories. The simulation results are then aggregated, including the parameter estimates and standard errors.

The summarized information is then used to calculate the average bias and coverage across each method, missingness proportion, and missingness mechanism. Bias is a measure of the error in the estimates caused by the imputation process. Measuring bias allows us to check that the imputation method does not cause the values to be consistently over- or underestimated. Using the parameters estimated from the original data set as the true values, the bias represents deviation from the ground truth. Bias is computed as follows:

$$\text{Bias}(\beta_k) = \frac{1}{N_{sim}} \sum_{i=1}^{N_{sim}} (\hat{\beta}_{k,i} - \beta_k),$$

where $\hat{\beta}_{k,i}$ is the estimate for β_k in the i-th iteration, β_k is the true parameter value, and N_{sim} is the total number of iterations.

Coverage measures the proportion of confidence intervals for each parameter that contain the true parameter value. Coverage reflects on the reliability of the methods in predicting values consistently close to the expected values. Coverage is computed using the following equation:

$$\text{Coverage} = \frac{1}{N_{sim}} \sum_{i=1}^{N_{sim}} I(\beta_k \in CI_{k,i}),$$

where N_{sim} is the total number of iterations, and $I(\beta_k \in CI_{k,i})$ is an indicator function that equals 1 if the confidence interval $CI_{k,i}$ contains the true parameter value β_k, and 0 otherwise.

Model standard error is a measure of the precision of an estimate. It measures the average standard error $\bar{SE}$ calculated for the parameter estimates. Lower model standard errors mean the model has less uncertainty in the point estimate $\hat{\beta}_k$. Model standard error is calculated as follows:

$$\bar{SE}(\hat{\beta}_k) = \frac{1}{N_{sim}} \sum_{i=1}^{N_{sim}} SE_i(\hat{\beta}_k),$$

where $\hat{\beta}_{k,i}$ is the estimated parameter with standard error $SE_i(\hat{\beta}_k)$ in the i-th iteration.

RESULTS

Simulation Results

The simulation results presented in this section are based on a MAR missingness mechanism for which our proposed imputation method and conventional multiple imputation are designed to work. Additional simulations considering a MCAR or a MNAR mechanism were performed and are presented in the appendix.

Figure 5 displays the average biases for each term across the missingness proportions within the MAR simulations. Bias is a measurement of the difference between the true value and the estimate produced by the model, indicating how often the model over- or underestimates the parameter. The figure illustrates how the ILR Regression method results in relatively low bias compared to the Scaling method. CCA also results in higher bias than the ILR regression method in most model terms, and especially in the intercept.

Figure 6 is a graph of the coverage probabilities for each method across missingness proportions. The coverage probability is the proportion of confidence intervals constructed using the data that contain the true parameter, which in this graphic is the true coefficient for each term in the model. Examining coverage probability allows us to evaluate the method's reliability because the coverage probability should match the nominal confidence interval level. From the figure, the ILR regression imputation consistently reaches the nominal level of 95% coverage at all missingness proportions, except in the ILR_2 variable where coverage is slightly decreased as the missingness proportion increases. The CCA method is unable to maintain high coverage in the ILR_2 variable with the coverage decreasing much more than the other methods. The Scaling method results in very low coverage that deviates somewhat from the expected coverage as the missingness proportion increases.

Figure 7 is a graph of the model standard error ($\bar{SE}(\hat{\beta}_k)$ for each method as the missingness proportion increases. Standard error is a measure of the uncertainty in the estimate from the fitted model, with a smaller SE indicating a more precise estimate. The figure shows that the standard errors for the ILR Regression and CCA methods increase at similar rates as the missingness proportion increases. The Scaling imputation method also results in increasing SE as the missingness proportion increases, though at a lesser degree than the CCA or ILR Regression methods. This is because as the missingness proportion increases and more observations are removed through the CCA process, the variability in the estimates increases as a result of the lower sample size.

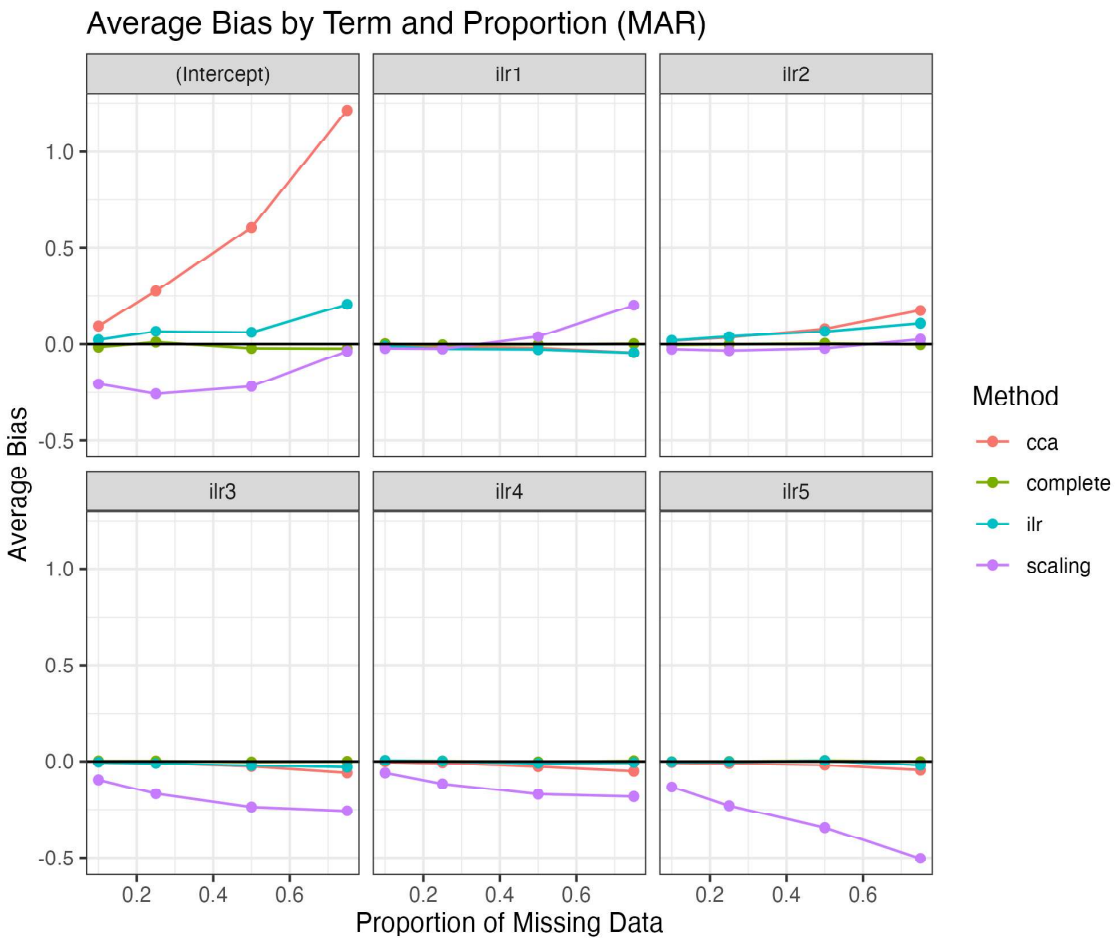

Figure 5. Graph of the average bias for each method at each missingness proportion. The averages were taken after computing the bias for each MAR simulation iteration.

Survey Data Analysis Results

One goal of the exercise motivations survey is to evaluate the relationship between the pie chart and EMI survey methods. Respondents were presented with 51 specific questions about their exercise motivations for the EMI method. The results were averaged into the six broader categories in order to match the six categories presented to the respondents for the pie chart method.

The novel ILR-transformation and scaling methods, as well as the existing CCA method, can be used to handle the missingness in the original exercise motivations dataset and draw conclusions about the relationship between the pie chart and EMI survey methods. **Figure 8** displays the missingness pattern present in the original dataset. There are three rows where each of the pie chart values are missing, including two rows where the missingness is across the entire observation. In addition to the two fully incomplete observations, one observation is missing the weight pie chart response, one is missing the social pie chart response, one is missing the challenge pie chart response, and one is missing the income demographic response.

The incomplete data were then analyzed three separate times using CCA, ILR regression imputation with MI, and scaling imputation with MI. The analysis model for each analysis was a multiple linear regression of the stress management EMI average on the predictors including the ILR transformed compositional survey responses and the respondent demographics. Stress management is used in Equation 4 as a single outcome to concisely illustrate the use of the methods, which can be readily extended to the other outcomes as well.

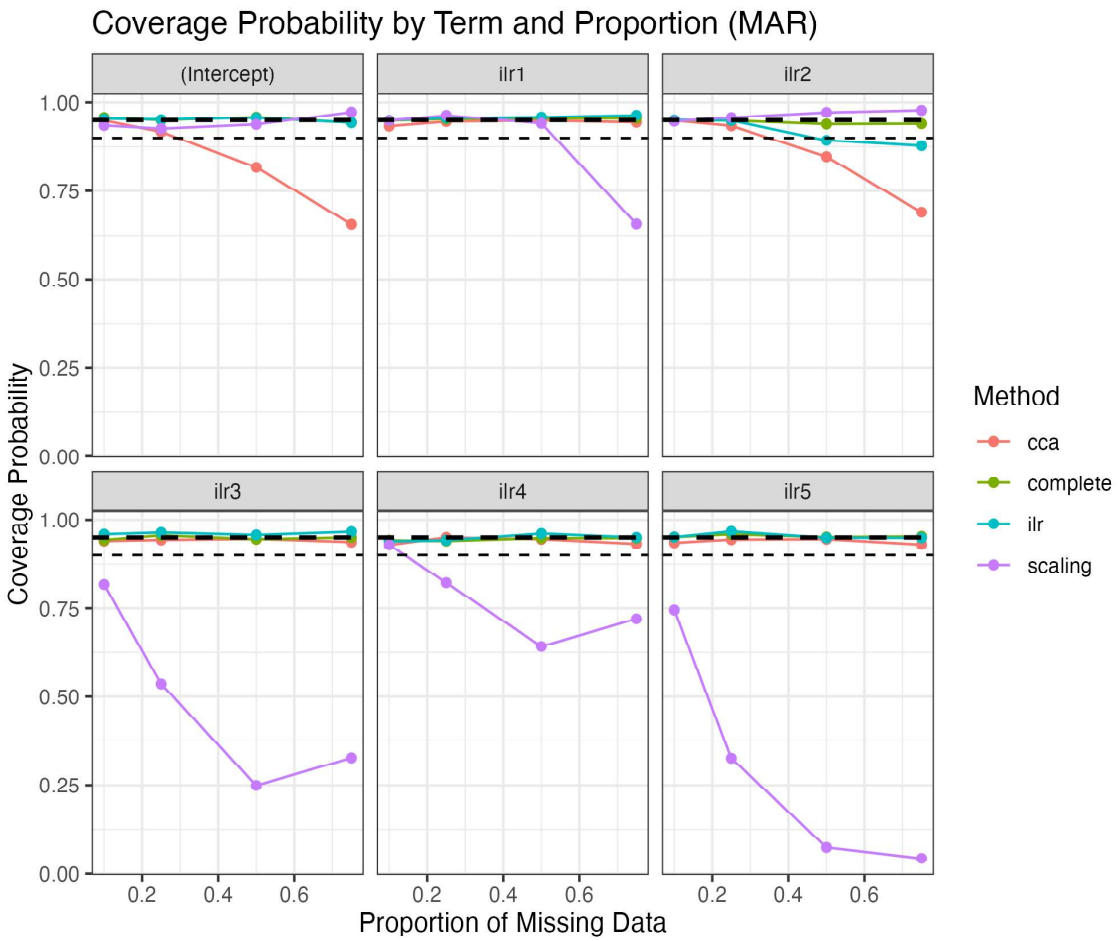

Figure 6. Graph of the coverage probability for each model term at each missingness proportion. The thinner line is at the 0.9 level, and the thicker line is at 0.95.

The estimates and their corresponding 95% confidence intervals are plotted in **Figure 9**, and **Table 2** contains the estimates, standard errors, and confidence intervals. The second ILR variable, having at most a High School diploma or GED, identifying as Female, earning between $40,000 and $49,999 annually, and identifying as Hispanic or Latino all have a significant impact on the importance of managing stress at a 95% confidence level. These significant variables are marked with an asterisk in **Table 2**.

DISCUSSION

When imputing compositional missing data, there are several considerations that need to be adequately handled to maintain the characteristics of the compositions and the legitimacy of the analysis. For example, maintaining the relationships between certain compositions is important so that imputed values are true to the structure and characteristics of the original data and respondents. Additionally, making sure that each observation sums to one is mathematically difficult to ensure. Finally, it is important to consider the mechanisms that cause the missingness, especially when it is related to experimental or social factors that are central to the experiment or study. For these reasons, it is essential that an imputation method for compositional data is well-rounded and applicable to different missingness mechanisms.

Through the simulation and following analysis, it appears that some of the evaluation metrics reflect better performance for the ILR Regression method than the Scaling method for the imputation of compositional missing data. The ILR Regression method results in consistently higher coverage probability than the Scaling method, as well as a lower magnitude of bias for all of the ILR model terms.

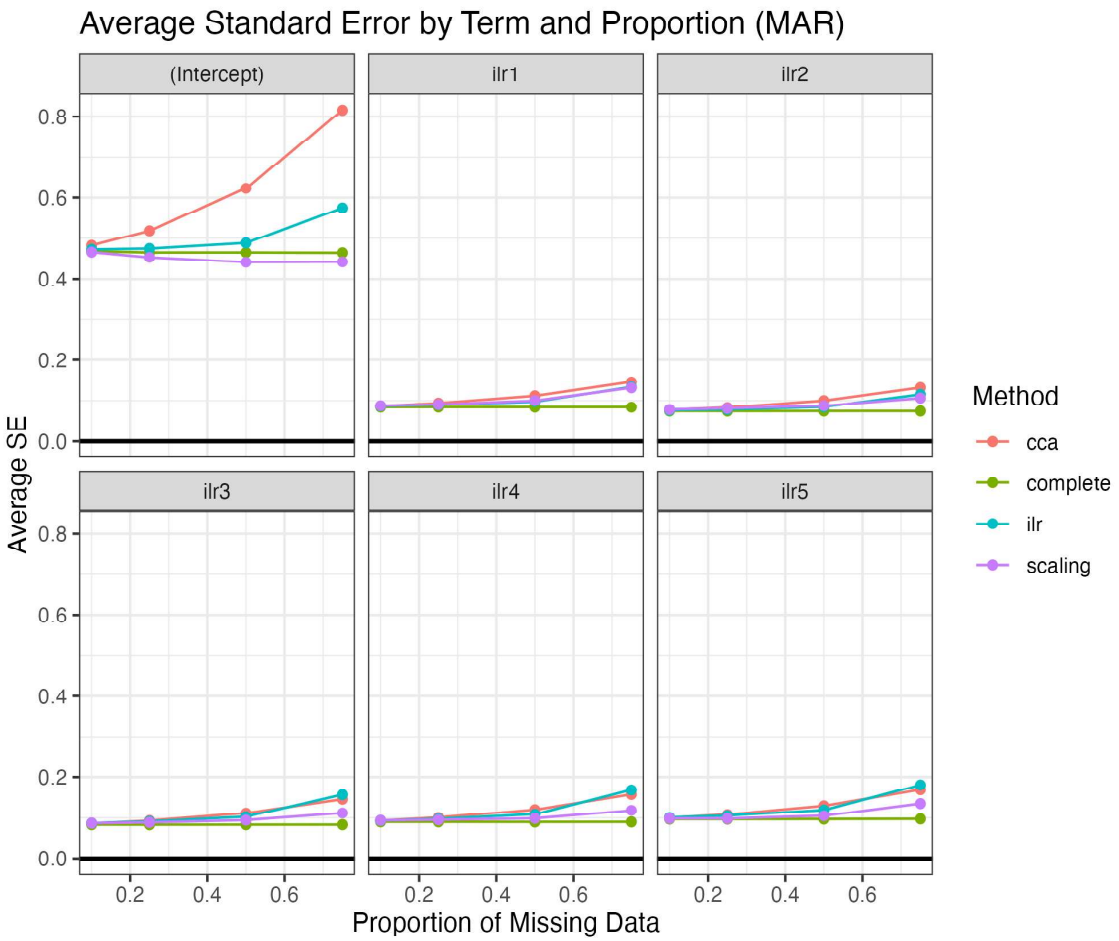

Figure 7. Graph of the standard error for each method at each missingness proportion.

To evaluate the performance of our methods under different missingness mechanisms, we also compared results across MCAR and MNAR scenarios. Results are reported graphically in the Appendix. Under MCAR, missingness is independent of both the outcome and the covariates. In contrast, the MNAR mechanism induced missingness on the second ILR variable depended directly on the values in the second ILR variable. Because of this, the MNAR results show that both Scaling and ILR Regression result in increased bias as the proportion of missingness increases, unlike the MAR and MCAR scenarios where only the Scaling method was affected. Similarly, in the MNAR case, coverage dips slightly for the second ILR variable when ILR regression is performed. These results make sense, given that MI is susceptible to increased bias in the MNAR scenario because the missingness is dependent on the values of the observed ILR values, which are a function of the pre-transformation pie-chart variables. In order to mitigate this bias when working with MNAR data, the dependency between the missingness and the observed ILR value should be captured in the imputation model.

The ILR Regression method results in consistently lower SE than the CCA method, and the magnitude of this difference increases at higher proportions of missingness. This indicates that CCA could be a reasonable missing data management method at low missingness proportions, but it results in significant uncertainty when more observations have missing data. Performing the two proposed imputation methods, alongside CCA, on the original dataset gives the opportunity to assess the performance of these methods as a sensitivity analysis when there is a very low proportion of missingness. From **Figure 9**, it appears that model estimates are somewhat similar across all approaches. However, some model terms (such as being in the $40,000-$49,999 income range or having at most a high school diploma) are significant when the dataset is imputed using the Scaling or ILR Regression methods, but insignificant when CCA is used.

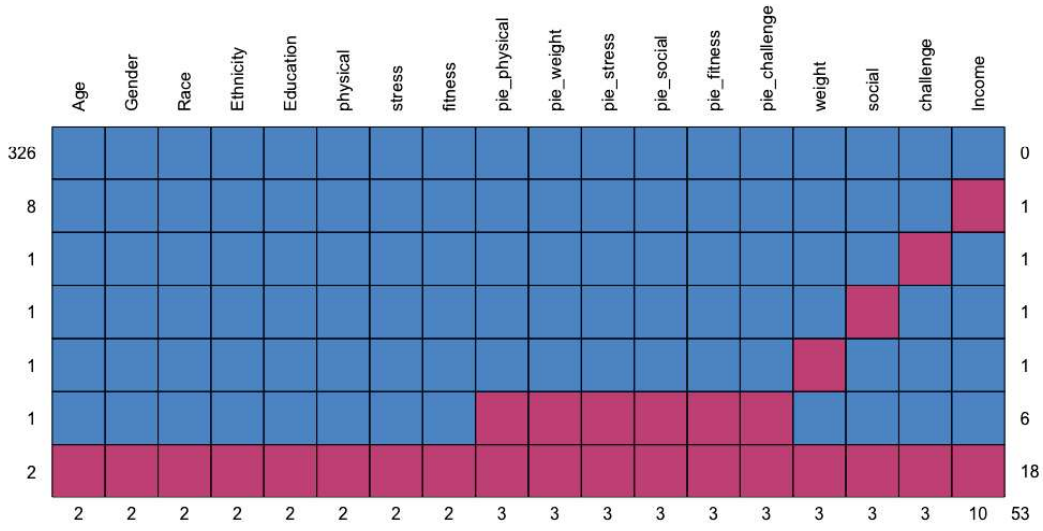

Figure 8. Missingness pattern in the original dataset.

Figure 9 also shows that CCA results in wider 95% confidence intervals around the estimates, which is a direct result of the smaller sample size associated with the higher level of missing observations. This makes a difference in the significance of some parameter estimates, and in applications where the proportion of missingness is much higher, this could make a more extreme impact on the bias of the model estimates.

CONCLUSION

When imputing incomplete compositional data, the imputation method must involve specific mathematical considerations to ensure that the imputed values maintain the characteristics of compositional data. These characteristics were considered in the development of the Scaling and ILR Regression imputation methods, and upon simulation analysis of these methods, we can see that the ILR Regression method can be used as an imputation model with multiple imputation. This method is a novel proposal for imputing missing compositional data, and it could be used in similar applications to effectively impute incomplete compositional data and use the completed datasets to generate a predictive model.

While the ILR method performs well for estimating EMI averages, interpreting the ILR variables' regression coefficients is difficult because ILR-transformed values do not have a direct relationship with the pre-ILR data. Therefore, it is not possible to use the ILR method in its current form to establish two-way relationships between the EMI averages and the compositional pie values. Future work could involve incorporating Principal Component Analysis (PCA) to find the relationships between each of the five ILR coefficients and the original columns. This would allow the impact of the ILR estimates on the EMI response variable to be measured in terms of the original compositional component categories, making the interpretation of the model more practically usable.

ACKNOWLEDGEMENTS

The authors thank the University of Connecticut Office of Undergraduate Research for funding this project through the IDEA Grant Program. The authors also thank Katherine Gnall for providing the exercise motivations dataset that

Term	CCA Estimate	CCA CI	ILR Estimate	ILR CI	Scaling Estimate	Scaling CI
Intercept	2.91	[1.06, 4.75]	3.40	[1.66, 5.14]	3.28	[1.54, 5.03]
Age	0.00	[-0.01, 0.01]	0.00	[-0.01, 0.01]	0.00	[-0.01, 0.01]
Not Cisgender	0.05	[-1.59, 1.69]	0.05	[-1.59, 1.69]	0.13	[-1.52, 1.77]
Less than High School	-0.99	[-2.53, 0.55]	-1.20	[-2.75, 0.34]	-1.12	[-2.64, 0.41]
High School/GED*	-0.21	[-0.60, 0.17]	-0.43	[-0.83, -0.03]	-0.48	[-0.87, -0.08]
Some College	N/A	[N/A, N/A]	-0.21	[-0.53, 0.10]	-0.17	[-0.48, 0.14]
Some Graduate School	-0.13	[-0.87, 0.61]	-0.34	[-1.09, 0.40]	-0.30	[-1.04, 0.45]
Graduate Degree	0.28	[-0.10, 0.65]	0.06	[-0.30, 0.42]	0.12	[-0.23, 0.48]
Hispanic or Latino*	-1.17	[-1.94, -0.39]	-1.17	[-1.94, -0.39]	-1.04	[-1.81, -0.27]
Ethnicity Unknown	0.34	[-0.20, 0.88]	0.34	[-0.20, 0.88]	0.26	[-0.28, 0.79]
Female*	-0.35	[-0.62, -0.09]	-0.35	[-0.62, -0.09]	-0.39	[-0.65, -0.13]
Non-Binary	-0.21	[-1.96, 1.54]	-0.21	[-1.96, 1.54]	-0.35	[-2.09, 1.40]
< $20,000	0.08	[-0.56, 0.73]	-0.15	[-0.59, 0.30]	-0.12	[-0.57, 0.32]
$20,000-$29,999	-0.05	[-0.77, 0.66]	-0.28	[-0.79, 0.23]	-0.29	[-0.80, 0.22]
$30,000-$39,999	0.28	[-0.41, 0.97]	0.05	[-0.45, 0.54]	0.04	[-0.47, 0.55]
$40,000-$49,999*	-0.36	[-1.02, 0.30]	-0.59	[-1.04, -0.14]	-0.56	[-1.01, -0.11]
$50,000-$59,999	0.19	[-0.50, 0.87]	-0.04	[-0.52, 0.43]	-0.06	[-0.54, 0.42]
$60,000-$69,999	0.09	[-0.58, 0.77]	-0.14	[-0.59, 0.31]	-0.16	[-0.61, 0.30]
$70,000-$79,999	0.10	[-0.62, 0.82]	-0.13	[-0.67, 0.40]	-0.14	[-0.68, 0.40]
$80,000-$89,999	0.20	[-0.61, 1.00]	-0.03	[-0.67, 0.60]	0.01	[-0.62, 0.65]
$90,000-$99,999	N/A	[N/A, N/A]	-0.23	[-0.86, 0.40]	-0.17	[-0.81, 0.47]
Black/African American	N/A	[N/A, N/A]	-0.04	[-0.40, 0.31]	-0.04	[-0.39, 0.32]
Native American	0.04	[-1.32, 1.39]	-0.01	[-1.33, 1.32]	0.14	[-1.21, 1.48]
Native Hawaiian	0.00	[-0.54, 0.55]	-0.04	[-0.51, 0.43]	-0.12	[-0.58, 0.34]
Asian	0.13	[-0.54, 0.79]	0.08	[-0.51, 0.68]	0.13	[-0.42, 0.68]
More than One Race	0.36	[-0.65, 1.37]	0.31	[-0.66, 1.29]	-0.01	[-0.94, 0.92]
ilr_pie_1	-0.11	[-0.28, 0.06]	-0.11	[-0.28, 0.06]	-0.10	[-0.27, 0.07]
ilr_pie_2*	0.38	[0.23, 0.53]	0.38	[0.23, 0.53]	0.36	[0.21, 0.51]
ilr_pie_3	-0.13	[-0.32, 0.06]	-0.13	[-0.32, 0.06]	-0.13	[-0.32, 0.05]
ilr_pie_4	-0.12	[-0.30, 0.06]	-0.12	[-0.30, 0.06]	-0.12	[-0.29, 0.06]
ilr_pie_5	-0.12	[-0.33, 0.09]	-0.12	[-0.33, 0.09]	-0.11	[-0.32, 0.10]

Table 2. Table of Estimates, Standard Errors, and 95% Confidence Intervals (CI) for CCA, ILR, and Scaling Methods. The base levels for the categorical variables are: Cisgender, College Graduate, Male, Not Hispanic or Latino, earning >$100,000, and identifying as White.

serves as the basis for this project. Our research was supported in part by an NSF AGEP-GRS supplement for Award #2015320.

The computational work performed on this project was done with help from the Storrs High-Performance Computing cluster. We thank the UConn Storrs HPC and HPC team for providing the resources and support that contributed to these results.

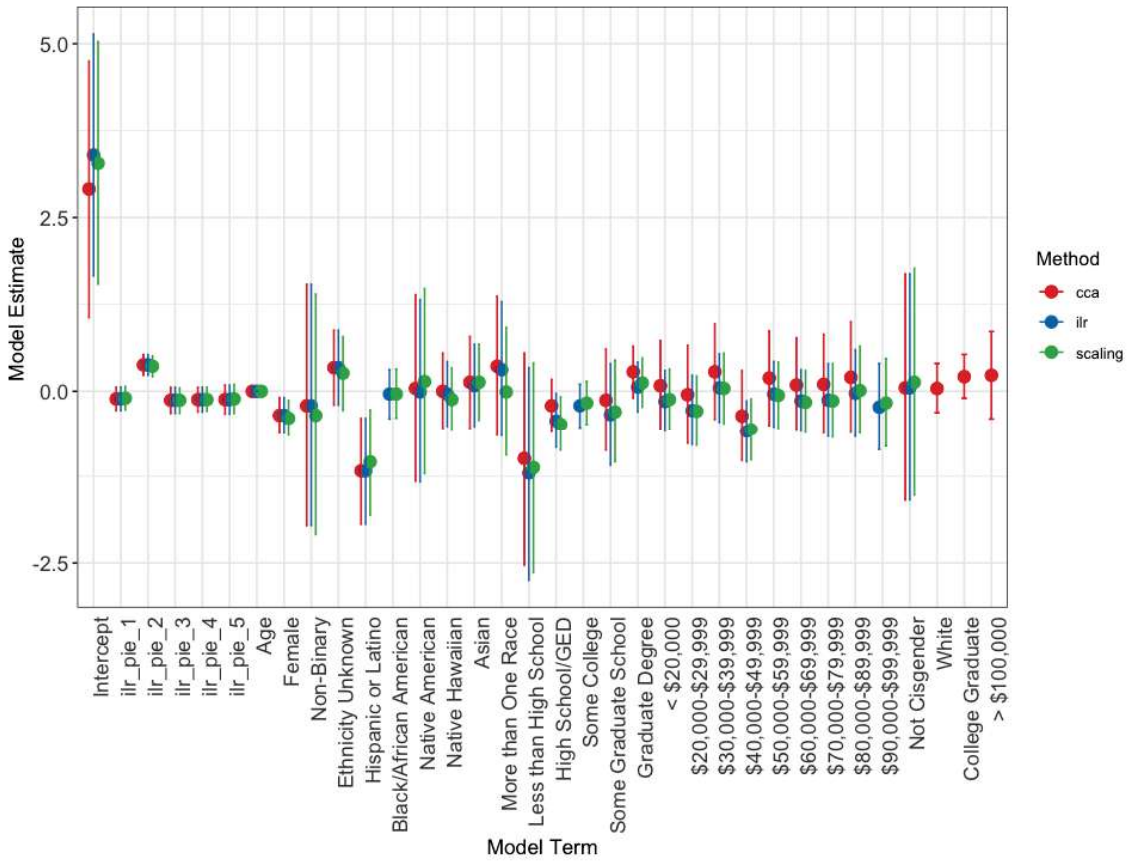

Figure 9. Dot plot with 95 percent confidence intervals representing the pooled parameter estimates, calculated after imputing the original (pre-simulation) dataset using each method.

REFERENCES

1. Ofer Harel and Xiao-Hua Zhou. Multiple imputation: Review of theory, implementation and software. *Statistics in Medicine*, 26(16):3057–3077, July 2007. ISSN 02776715, 10970258. doi: 10.1002/sim.2787.
2. Roderick Little and Donald Rubin. *Statistical Analysis with Missing Data*. John Wiley & Sons, 04 2019. ISBN 9780470526798. doi: 10.1002/9781119482260.
3. Donald B. Rubin. *Multiple Imputation for Nonresponse in Surveys*. John Wiley & Sons, Ltd, 1st edition, 1987. doi: 10.1002/9780470316696.
4. Tim P. Morris, Ian R. White, and Patrick Royston. Tuning multiple imputation by predictive mean matching and local residual draws. *BMC medical research methodology*, 14:75, 06 2014. doi: 10.1186/1471-2288-14-75.
5. L.L. Doove, S. van Buuren, and E. Dusseldorp. Recursive partitioning for missing data imputation in the presence of interaction effects. *Computational Statistics & Data Analysis*, 72:92–104, 2014. ISSN 0167-9473. doi: https://doi.org/10.1016/j.csda.2013.10.025. URL *https://www.sciencedirect.com/science/article/pii/S0167947313003939*.
6. John Aitchison. Principles of compositional data analysis. *Lecture Notes-Monograph Series*, 24:73–81, 1994. ISSN 0749-2170. URL *http://www.jstor.org/stable/4355794*.
7. V. Pawlowsky-Glahn and J. J. Egozcue. Geometric approach to statistical analysis on the simplex. *Stochastic Environmental Research and Risk Assessment*, 15(5):384–398, October 2001. ISSN 1436-3259. doi: 10.1007/s004770100077.
8. Donald B. Rubin. Multiple imputation after 18+ years. *Journal of the American Statistical Association*, 91(434):473–489, 1996. ISSN 01621459. doi: 10.1080/01621459.1996.10476908. URL *http://www.jstor.org/stable/2291635*.
9. Joseph L. Schafer and John W. Graham. Missing data: Our view of the state of the art. *Psychological Methods*, 7(2): 147–177, June 2002. ISSN 1082-989X. doi: 10.1037/1082-989X.7.2.147.

10. Donald B. Rubin. Inference and missing data. *Biometrika*, 63(3):581–592, 1976. ISSN 00063444, 14643510. doi: 10.1093/biomet/63.3.581. URL *http://www.jstor.org/stable/2335739*.
11. Trivellore E Raghunathan, James M Lepkowski, John Van Hoewyk, and Peter Solenberger. A Multivariate Technique for Multiply Imputing Missing Values Using a Sequence of Regression Models. *Statistics Canada*, 27(1):85–95, June 2001.
12. Michael G Kenward and James Carpenter. Multiple imputation: Current perspectives. *Statistical Methods in Medical Research*, 16(3):199–218, June 2007. ISSN 0962-2802. doi: 10.1177/0962280206075304.
13. R Core Team. *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing, Vienna, Austria, 2021. URL *https://www.R-project.org/*.
14. Stef van Buuren and Karin Groothuis-Oudshoorn. Mice: Multivariate Imputation by Chained Equations in R. *Journal of Statistical Software*, 45:1–67, December 2011. ISSN 1548-7660. doi: 10.18637/jss.v045.i03.
15. J. Aitchison. The statistical analysis of compositional data. *Journal of the Royal Statistical Society. Series B (Methodological)*, 44(2):139–177, 1982. ISSN 00359246. URL *http://www.jstor.org/stable/2345821*.
16. K. Gerald van den Boogaart, Raimon Tolosana-Delgado, and Matevz Bren. *compositions: Compositional Data Analysis*, 2023. URL *https://CRAN.R-project.org/package=compositions*. R package version 2.0-6.
17. David Markland. The Exercise Motivations Inventory, 1997.
18. Stef van Buuren. *Flexible Imputation of Missing Data*. Interdisciplinary Statistics Series. Chapman and Hall/CRC, 2nd edition, 2018. URL *https://stefvanbuuren.name/fimd/*.
19. J Aitchison. *The statistical analysis of compositional data*. Chapman & Hall, Ltd., GBR, 1986. ISBN 0412280604.
20. Rachael A. Hughes, Ian R. White, Shaun R. Seaman, James R. Carpenter, Kate Tilling, and Jonathan AC Sterne. Joint modelling rationale for chained equations. *BMC Medical Research Methodology*, 14(1):28, February 2014. ISSN 1471-2288. doi: 10.1186/1471-2288-14-28.
21. Tim P. Morris, Ian R. White, and Michael J. Crowther. Using simulation studies to evaluate statistical methods. *Statistics in Medicine*, 38(11):2074–2102, 2019. doi: 10.1002/sim.8086.
22. Hanne I. Oberman and Gerko Vink. Toward a standardized evaluation of imputation methodology. *Biometrical Journal*, n/a(n/a):2200107, 2024. ISSN 1521-4036. doi: 10.1002/bimj.202200107.

ABOUT THE STUDENT AUTHOR

Sana Gupta graduated from the University of Connecticut with a B.S. in Statistics in Spring of 2024. She is now a Second-Year Statistics Ph.D. student at the University of Connecticut.

PRESS SUMMARY

In this paper, we present a novel method for imputing missing values in surveys involving compositional data. Unlike traditional techniques, our approach involves the use of the Isometric Log-Ratio transformation to maintain the proportional nature of the data. The method was evaluated by applying it to a dataset of responses from an exercise motivations study and the results were compared with standard methods. The ILR-based method resulted in consistent performance at different missingness proportions.