Variable Selection on Compositional Data based on Variable Deletion

David Perez-Suarez

M.A., University of North Carolina at Greensboro, United States, 2018
B.A., University of North Carolina at Greensboro, United States, 2017

A Dissertation Presented to the Graduate Faculty
of University of Virginia in Candidacy for the Degree of
Doctor of Philosophy

Department of Statistics

University of Virginia
November, 2023

Abstract

Compositional data consists of proportions or percentages of compositions, which are usually positive vectors, with the relevant information being the ratios between their components (Egozcue and Pawlowsky-Glahn (2006)). The unique feature of compositional data is that the observed values of compositional variables sum to 1 for each subject, and this feature makes the selection for informative variables challenging when dimensionality is high since many of the existing variable selection methods cannot accommodate this data structure. Compositional data appears in a wide range of applications such as geology, consumer demand analysis, forensic science, etc., and an effective variable selection method for such data is highly desired. In this work, we developed a variable selection method for compositional data in a linear regression model. The developed method is based on the deletion of the subsets of the variables and the corresponding changes in the coefficient of determination. The deletion method was computed efficiently. The numerical performance of the developed method is satisfactory in simulation studies. This variable selection method for compositional data can also be generalized for more complicated models.

Acknowledgements

I want to thank everyone who helped and encouraged me throughout the completion of this thesis.

I want to thank Prof. Jianhui Zhou for being grateful enough to be my advisor throughout this thesis. He always helped out whenever I was stuck in a complex problem or concept. He always encouraged me to start things early, especially this thesis, so I did not panic or stress out at the last minute. Furthermore, he was always approachable as he was easy to talk to and gave me great advice on both this thesis and in life, whether that was in job applications, internships, etc. My PhD study would not have run smoothly without his guidance and support.

I am grateful to Prof. Jennie Ma for providing me with the necessary dataset to complete the data analysis portion of my thesis. She was very kind in letting me know what the dataset was all about, what to focus on primarily, and any questions I had about it. Without her, I would not have completed my thesis promptly.

I would also like to thank Prof. Jordan Rodu, Prof. Cynthia Tong, and Prof. Shan Yu. They were my committee members throughout this thesis. Their expertise in statistics, kindness, and advice helped make my thesis more efficient than before and more accessible to complete.

Finally, I would like to thank my parents, especially my father, for their love and support. Without their advice and encouragement, I would not have

made it this far into this thesis.

Contents

Chapter 1

Introduction to Compositional data

According to Greenacre (2019), compositional data are samples of non-negative multivariate data that have been expressed relative to a fixed total, usually as propositions or compositions summing to 1 or percentages summing to 100%. In Egozcue and Pawlowsky-Glahn (2006), compositions are defined as positive vectors with relevant information in the ratios between their components. The idea behind it is as follows (Aitchison and Shen (1980)).

Let $\mathbb{R}^p$ denote p-dimensional real space and $\mathbb{S}^{p-1}$ its positive simplex such that it is represented as

$$\mathbb{S}^{p-1} = \{(x_1, ..., x_p) : x_i > 0,\ i = 1, ..., p,\ \sum_{i=1}^{p} x_i = 1\}.$$

The symbol $\mathbf{x}$ and $\mathbf{y}$ are reserved for vectors in $\mathbb{S}^{p-1}$ and $\mathbb{R}^p$, respectively. Any vector or point $\mathbf{x}$ in $\mathbb{S}^{p-1}$ is termed a composition and any collection of such vectors, compositional data.

The simplex plays a vital role as a sample space in many practical situations where compositional data analysis results, in the form of proportions of some whole, require interpretation. Statistical analysis of such data has proved difficult because of a lack of both concepts of independence and rich enough parametric classes of distributions in the simplex (Aitchison (1982)). Hence, Aitchison (1982) came up with a variety of interrelated independence hypotheses and new classes of transformed-normal distributions in the simplex that are provided where the independence hypotheses can be tested through the standard theory of parametric hypothesis testing. These new concepts and statistical methodology are illustrated by a number of applications such as geology and consumer demand analysis.

Geological literature has many problems interpreting rock and sediment specimens' chemical, mineral, and fossil compositions. Each composition of each specimen is a set of 3 to 20 proportions that sum to unity and thus can be represented by a point in an appropriately dimensional simplex. Hence, compositional data is used to find classes of parametric models for describing the experienced pattern of variability, investigating the adequacy of such models, and testing several independence hypotheses for the pertinent sets of proportions.

An essential aspect of the study of consumer demand is the analysis of

household budget surveys. Attention is focused on several expenditures on mutually exclusive and exhaustive commodity groups and their relations to total expenditure, income, household compositions, etc. Hence, compositional data is used to determine the role of the composition of expenditures and the proportion of total expenditure allocated to such commodity groups in a form of a budget-share approach to the analysis.

There are a lot of practical problems where the simplex $\mathbb{S}^{p-1}$ forms the whole, or even a significant part, of the sample space. For such problems, concepts of independence play a key role (Aitchison (1982)). However, the simplex has proved too awkward to handle statistically due to the difficulties appearing in the scarcity of both meaningful definitions of independence and measures of dependence and in the absence of satisfactory parametric classes of distributions on $\mathbb{S}^{p-1}$ (Aitchison (1982)). Also, due to the unit-sum constraint, the p components of a composition cannot vary freely (Lin et al., 2014). Hence, traditional methodology often requires the omission of specific components to ensure identifiability and, thus, results in intrinsic difficulties in proving meaningful interpretations for the regression parameters. Since the seminal work of Aitchison (1982), methodological developments for compositional data analysis have given rise to groundbreaking and effective research to deal with such problems in linear regression, principal components analysis, and missing data.

Beginning with Aitchison and Bacon-Shone (1984), much research was devoted to finding a useful transformation for compositional data in the context

of principal component analysis (PCA). The centred logratio (clr) transformation turned out to be a preferable option (Aitchison and Greenacre (2002)) which is defined as

$$clr(x) = [ln(\frac{x_1}{g(x)}), ln(\frac{x_2}{g(x)}), ..., ln(\frac{x_p}{g(x)})] = ln(\frac{x}{g(x)})$$

with $x = (x_1, x_2, ..., x_p)$ such that $x \in \mathbb{S}^p$ and $g(x) = \sqrt{\prod_{i=1}^{p} x_i}$ is the geometric mean of the compositions of x. It is based on dividing each sample by the geometric mean of its values and taking the logarithm. The principal components (PCs) are then aimed at summarizing the multivariate data structure and subsequently can be used for dimension reduction. The goal of keeping the most important data information with only a few PCs can fail for data containing outliers because these can spoil the estimation of the PCs (Maronna et al. (2019)). This artifact arises for classical PCA, where the estimation of the PCs is based on the classical sample covariance matrix. As a solution, robust PCA uses a robust estimation of the covariance matrix, and the PCs will still point in directions of the main variability of the majority of data (Filzmoser (1999)). However, this procedure does not work with clr transformed data due to robust covariance estimators usually needing a full rank data matrix. Hence, Filzmoser et al. (2009) solved this problem by taking the isometric logratio transformation (ilr) instead, where

the ilr transformation is defined as

$$ilr(x) = V^t clr(x) \tag{1.1}$$

with $clr(x)$ being the centred log ratio transformation of x and $V \in \mathbb{S}^{p-1}$ is a matrix whose columns form an orthonormal basis of the clr-plane. However, the ilr transformation has the disadvantage that the resulting new variables are no longer directly interpretable in terms of the originally entered variables. Hence, Filzmoser et al. (2009) proposed a technique on how a robust PCA's resulting scores and loadings on ilr transformed data can be back-transformed and interpreted. Their procedure is demonstrated using a real data set from regional geochemistry and compared to results from non-transformed and non-robust versions of PCA. Their procedure using ilr-transformed data and robust PCA yielded superior results to all other approaches. Hence, the examples show that due to the compositional nature of geochemical data, PCA should not be carried out without an appropriate transformation.

Multivariate statistics methods, including regression analysis, have been adopted to model compositional data, but the existing research is still scattered and fragmented (Wang et al. (2013)). Hence, Wang et al. (2013) contributed to the linear regression modeling for compositional data, which provided an innovative way for parameter estimation, model evaluation, and interpretation. From the modeling viewpoint, operators of compositional

data vectors, such as perturbation operation, power transformation, and inner product, are proposed in Aitchison geometry. Also, Wang et al. (2013) established a regression model in orthonormal coordinates after transforming the compositional data into real vectors by the ilr transformation. It turned out that these two modeling methods are entirely equivalent in essence and highly efficient, as evidenced by the results from Wang et al. (2013). Hence, their inner product definition for compositional data is reasonable.

When working with actual data, using values lower than the maximum limit will not introduce much distortion in the data structure (Real et al. (2011)). However, if the data must be transformed into logarithms before the analysis, the effects on the data structure can be significant. If the imputed values are between 0 and 1, the smaller the value, the larger the negative value from transformation. The smallest possible value, zero, is useless due to its undefined logarithm. As mentioned, data adding to a fixed constant like one, like proportions, are known as compositional data. The methods designed for their analysis are based on logarithms of ratios among the variables (Real et al. (2011)). Hence, compositional data are susceptible to the problem of missing values, and imputing these values is a delicate procedure. Applying standard statistical methods like correlation analysis, imputation methods, or principal component analysis directly to compositional data would give misleading results. Hron et al. (2010) introduced a new imputation algorithm for missing values in compositional data by proposing an iterative model-based imputation technique. The method is based on iterative regressions,

accounting for multivariate data information. The regressions must be performed in a transformed space, and classical or robust regression techniques can be employed depending on the data quality. Their proposed method is tested on real and simulated data sets; the results show that it outperforms standard imputation methods. In the presence of outliers, the model-based method with robust regression is preferred.

1.1 Review of variable selection in linear regression

Assume that we have p predictor variables $X_1, X_2, ..., X_p$. Then, a multiple linear regression model takes the form

$$Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \ldots + \beta_p X_p + \epsilon \tag{1.2}$$

where Y is the response variable, X_j is the jth predictor variable, β_j is the average effect on Y of a one unit increase in X_j, holding all other predictors fixed, and ϵ is the error term. The values for $\beta_0, \beta_1, \beta_2, ..., \beta_p$ are estimated using the least squares method, which minimizes the sum of squared residuals (RSS)

$$RSS = \sum_{i=1}^{n} (y_i - (\beta_0 + ... + \beta_p x_{ip}))^2$$

where y_i is the observed response value for the ith observation. For such a model, we are interested in performing variable selection to choose the pertinent subset of X that has an effect on Y. Some variable selection methods that are used in multiple linear regression include but are not limited to forward, backward, stepwise, and penalized (Lasso, etc.) selection.

1.1.1 Forward selection

In this method, we build a regression model from a set of predictor variables by entering predictors until there is no statistically valid reason to enter anymore. The steps for this procedure are as follows.

1. Let M_0 denote the null model, which has no predictor variables.
2. For $k = 0, 2, ..., p - 1$:
    - Fit all $p - k$ models that increase the predictors in M_k with one additional predictor variable.
    - Choose the best among these $p-k$ models and call it M_{k+1}. Define the "best" model as that with the highest R^2 or lowest RSS.
3. Select a single best model from among $M_0, M_1, ..., M_p$ using the lowest cross-validation prediction error, lowest C_p, lowest BIC, lowest AIC, or highest adjusted R^2

1.1.2 Backward selection

In this method, we build a regression model from a set of predictor variables by removing predictors until there is no statistically valid reason to remove any more. The steps for this procedure are as follows.

1. Let M_p denote the full model, which has all p predictor variables
2. For $k = p, p-1, ..., 1$:
    - Fit all k models that contain all but one of the predictors in M_k, for a total of $k-1$ predictor variables.
    - Choose the best among these k models and call it M_{k-1}. Define the "best" model as that with the highest R^2 or lowest RSS
3. Select a single best model from among $M_0, M_1, ..., M_p$ using the lowest cross-validation prediction error, lowest C_p, lowest BIC, lowest AIC, or highest adjusted R^2

1.2 Stepwise selection

We can use this procedure to build a regression model from a set of predictor variables by entering and removing predictors stepwise into the model until there is no statistically valid reason to enter or remove any more. It combines both forward and backward selection to choose the "best" model by the end. The steps for this procedure are as follows.

1. Fit an intercept-only model

2. Add predictors to the model sequentially, just like in forward selection. However, after adding each predictor, we removed any predictors that no longer improved the model, just like in backward selection.

3. Repeat Steps 1-2 until a single best model is reached by way of the lowest cross-validation prediction error, lowest C_p, lowest BIC, lowest AIC, or highest adjusted R^2

One benefit of this procedure is that it is more computationally efficient than best subset selection below. Given p predictor variables, the best subset selection must fit 2^p models, while stepwise selection only has to fit $1 + \frac{p(p+1)}{2}$ models. For instance, for $p = 10$ predictor variables, best subset selection must fit 1024 models, while stepwise selection only fits 56. Furthermore, one drawback to the stepwise procedure is that it is not guaranteed to find the best possible model out of all $1 + \frac{p(p+1)}{2}$ potential models. For instance, let there be a dataset with $p = 4$ predictors. The best possible one-predictor model may contain x_1, and the best possible two-predictor model may instead contain x_1 and x_3. Hence, stepwise selection will fail to select the best possible two-predictor model because M_1 will contain x_1, so M_2 must also contain x_1 along with some other variable.

1.3 Penalized regression

In general, the optimization problem in penalized linear regression is given by

$$\begin{aligned}\hat{\beta}_{\text{pen}} &= \underset{\beta}{\text{argmin}}(\|y - X\beta\|^2 + \lambda\phi(\beta)) \\ &= \underset{\beta}{\text{argmin}}((y_i - \beta_0 - \sum_{j=1}^{p} x_{ij}\beta_j)^2 + \lambda\phi(\beta_1, ..., \beta_p))\end{aligned}$$

where the parameter $\lambda \geq 0$ is a tuning parameter that controls the shrinkage amount, the intercept β_0 is generally not shrunken, and the function ϕ is the penalty function on β.

1.3.1 Lasso

When $\phi(\beta) = \|\beta\|_1 = \sum_{j=1}^{p} |\beta_j|$, then the penalized regression becomes the least absolute shrinkage and selection operator (lasso). The optimization problem in lasso is given by

$$\begin{aligned}\hat{\beta}_{\text{Lasso}} &= \underset{\beta}{\text{argmin}}(\|y - X\beta\|_2^2 + \lambda\|\beta\|_1) \\ &= \underset{\beta}{\text{argmin}}((y_i - \beta_0 - \sum_{j=1}^{p} x_{ij}\beta_j)^2 \text{ subject to } \sum_{j=1}^{p} |\beta_j| \leq t.\end{aligned}$$

Like before, the parameter $\lambda \geq 0$ is the tuning parameter that controls the amount of shrinkage, and the intercept β_0 is not shrunken. If t is sufficiently small, then some of the coefficients will be exactly zero whereas if t is larger

than $t_0 = \sum_{j=1}^{p} |\hat{\beta}_j|$, then $\hat{\beta}_{\text{Lasso}}$ are the OLS estimates $\hat{\beta}_j$. Note that $\hat{\beta}_{\text{Lasso}}$ cannot be given in closed form, but efficient algorithms, such as coordinate descent, are used to compute the entire path of solutions as λ is varied. The coordinate descent algorithm is described in details in the next paragraph.

For any fixed λ, given the current value $\tilde{\beta}(t)$ in the tth iteration. For t=0,1,..., the algorithm for determining $\hat{\beta}$ is

1. Calculate

$$\tilde{z}_j = n^{-1} \sum_{i=1}^{n} x_{ij} r_i + \tilde{\beta}_j^{(s)}$$

   where $r_i = y_i - \tilde{y}_i$ is the current residual

2. Update $\tilde{\beta}_j^{(s+1)}$ using $\tilde{\beta}_j = S(\tilde{z}_j; \lambda)$ such that

$$S(z; \lambda) = \begin{cases} z - \lambda, & z > \lambda \\ 0, & |z| < \lambda \\ z + \lambda, & z < -\lambda \end{cases}$$

3. Update $r_i \to r_i - (\beta_j^{(s+1)} - \beta_j^{(s)}) x_{ij}$ for all i

One benefit that the lasso has lies in the bias-variance tradeoff. Recall that the MSE (mean-squared error) is used to measure the accuracy of a given model and is calculated as MSE=Variance+Bias^2+irreducible error. With lasso, we want to introduce a slight bias so that the variance can be

substantially decreased, leading to a lower MSE overall. Furthermore, one drawback to this method is that it tends to omit covariates with small coefficients (Bühlmann et al. (2011)). This problem arises because lasso minimizes prediction error subject to the constraint that the model is not too complex, and lasso measures complexity by the sum of the absolute values of the coefficients. Covariates with small coefficients tend to be entrapped by the constraint. Small coefficients of covariates that belong in the model look like small coefficients of variables that do not. That bias is not solely a function of the coefficient's size.

1.3.2 Other Penalized Regression methods

Remember that one can use other penalized regression methods to achieve the desired result. Some of these methods include but are not limited to, the adaptive lasso, smoothly clipped absolute deviation (SCAD), and minimax concave penalty (MCP). For adaptive lasso, the penalty function is $\phi(\beta) = \sum_{j=1}^{p} w_j|\beta_j|$ such that $w_j = |\tilde{\beta}_j|^{-1}$. For this method, note that smaller weights lead to larger $\hat{\beta}$ whereas $w_j = \infty$ leads to $\hat{\beta}_j = 0$. Furthermore, when $\lambda\phi(\beta) = \sum_{j=1}^{p} P(\beta_j|\lambda, \gamma)$ such that

$$P(x|\lambda,\gamma) = \begin{cases} \lambda|x|, & |x| \leq \lambda \\ \frac{2\gamma\lambda|x| - x^2 - \lambda^2}{2(\gamma-1)}, & \lambda < |x| < \gamma\lambda \\ \frac{\lambda^2(\gamma+1)}{2}, & |x| \geq \gamma\lambda, \end{cases}$$

it is the smoothly clipped absolute deviations (SCAD) method. For such a method, note that SCAD first coincides with the lasso until $|x| = \lambda$, then smoothly transitions to a quadratic function until $|x| = \gamma\lambda$, after which it remains constant for all $|x| > \gamma\lambda$. Lastly, the minimax concave penalty (MCP) specifies the penalty function $\lambda\phi(\beta) = \sum_{j=1}^{p} P(\beta_j|\lambda, \gamma)$ such that

$$P(x|\lambda,\gamma) = \begin{cases} \lambda|x| - \frac{x^2}{2\gamma}, & |x| \leq \gamma\lambda \\ \frac{1}{2}\gamma\lambda^2, & |x| \geq \gamma\lambda. \end{cases}$$

For such a method, just like the SCAD, it starts by applying the same rate of penalization as the lasso, then smoothly relaxes the rate down to zero as the absolute coefficient value increases. Compared to SCAD, MCP relaxes the penalization rate automatically, while with SCAD, the rate remains constant for a while before decreasing.

1.4 Review of variable selection in compositional data

Variable selection is one particular area for high dimensional compositional data to alleviate the collinearity. Recent works, such as Hron et al. (2013), Lin et al. (2014) and Susin et al. (2020), proposed methods to perform variable selection for compositional data.

Most statistical methods are designed for the usual Euclidean geometry,

and hence, compositional data first needs to be transformed from the simplex to the real space. Hron et al. (2013) considered the so-called log-ratio transformation for such a purpose. The centred logratio (clr) transformation is an isometric mapping between S^p and a hyperplane of $\mathbb{R}^p$. Due to its construction, the clr variables lead to collinear data due to $\sum_{i=1}^{p} y_i = 0$. This had consequences if the statistical methods required data with full rank, which was required for computing an inverse covariance matrix. Despite this dilemma, clr variables are still frequently used due to their intuitive interpretation and relation to a particular choice of the ilr transformation.

The ilr transformation represents an isometric mapping from S^p to $\mathbb{R}^{p-1}$, and has an additional advantageous feature in that it represents compositions in coordinates of an orthonormal basis on the simplex. On the contrary, the clr transformation results in coordinates with respect to a generating system (note that the dimension of the simplex equals $p-1$). Consequently, the resulting ilr data matrix has full rank, and possible numerical problems may be avoided. There are many possible ways to construct the ilr coordinates. A pertinent choice was to use sequential binary partition (Egozcue and Pawlowsky-Glahn (2005)), wherein each of the $p-1$ steps of the procedure the compositional parts are divided into two nonoverlapping groups; the resulting $p-1$ ilr variables represent balances between these groups. An alternative interpretation of ilr coordinates is based on a decomposition of their covariance structure (Fišerová et al. (2011)): each balance explains log-ratios between compositional parts in both groups that come from the

corresponding step of a sequential binary partition.

Consequently, the resulting balances uniquely represent all log ratios in the composition. Furthermore, properly choosing the balances makes it possible to proceed from the full composition to a subcomposition, which is a subset of a composition. Thus, a stepwise procedure was derived to obtain such a subcomposition, where the effect of the change of the information that had the resulting subcomposition is rather negligible. More details on how this procedure was constructed will be explained in the following paragraphs.

Consider a composition $x = (x_1, ..., x_p)'$ and $y_i = \sqrt{\frac{p-1}{p}} z_1^{(i)}, i = 1, ..., p$ such that $z_i^{(l)} = \sqrt{\frac{p-i}{p-i+1}} ln(\frac{x_i^{(l)}}{\sqrt[p-i]{\prod_{j=i+1}^{p} x_j^{(l)}}}), i = 1, ..., p-1, l = 1, ..., p$. Without loss of generality, let

$$var(y_1) \geq ... \geq var(y_p)$$

which is equivalent to

$$\sum_{k=1}^{p} var(ln(\frac{x_1}{x_k})) \geq \sum_{k=1}^{p} var(ln(\frac{x_2}{x_k})) \geq ... \geq \sum_{k=1}^{p} var(ln(\frac{x_p}{x_k})).$$

Since y_p has the smallest variance, its contribution to the compositional data set's overall variance, $totvar(x)$, is defined as

$$totvar(x) = \frac{1}{2p} \sum_{i=1}^{p} \sum_{j=1}^{p} var(ln(\frac{x_i}{x_j})),$$

is minimal. This is equivalent to the statement that the log-ratios' aggre-

gated variances with the part x_p have the smallest contribution to the overall variance. Consequently, the part x_p does not determine the multivariate data structure, and it can be omitted from the composition. Hence, we have the subcomposition $\mathbf{x}_1 = (x_1, ..., x_{p-1})'$. In the next step, we perform a clr transformation on $\mathbf{x}_1$, calculating the variances of the clr transformed variables and omitting the part corresponding to the clr variable with the smallest variance. Then, we continue until a certain number of parts is obtained, and we stop at the latest after $p - 2$ steps.

Practitioners are often interested in reducing the number of compositional variables for statistical analysis because this simplifies the analysis and the interpretation of results (Hron et al. (2013)). An intuitive selection of variables based on expert knowledge of subject matter specialists may lead to significant changes in the multivariate statistical analysis results. For example, experts may be interested in analyzing specific geochemical processes and selecting elements for the statistical analysis that are somehow related to these processes. In this selection, they may miss variables responsible for substantial information to the multivariate information, and their omission changes the statement about the resulting subcomposition. Note that this selection is against the general definition of compositional data as multivariate observations where the only relevant information is contained in the ratios between the parts (Egozcue and Pawlowsky-Glahn (2006)).

Regarding compositional data, there are intrinsic difficulties in providing sensible interpretations for the regression parameters. To solve such diffi-

culties, Aitchison and Bacon-Shone (1984) considered applying the log-ratio transformation (Aitchison (1982)) to compositional covariates, resulting in the linear log-contrast model.

$$y = Z^p \beta^*_{-p} + \epsilon, \tag{1.3}$$

where $Z^p = \{log(\frac{x_{ij}}{x_{ip}})\}$ is the n x $(p-1)$ log-ratio matrix whose pth component is the reference component, $\beta^*_{-p} = (\beta^*_1, ..., \beta^*_{p-1})$ is the corresponding $(p-1)$-vector of regression coefficients, and ϵ is a n-vector of independent noise distributed as $N(0, \sigma^2)$. By having a new coefficient $\beta^*_p = -\sum_{j=1}^{p-1} \beta^*_j$, model (1.3) can be expressed as

$$y = Z\beta^* + \epsilon, \text{ subject to } \sum_{j=1}^{p} \beta^*_j = 0 \tag{1.4}$$

where $Z = (z_1, ..., z_p) = (log(x_{ij}))$ is the n x p design matrix and $\beta^* = (\beta^*_1, ..., \beta^*_p)^T$ is the p-vector of regression coefficients.

Applying the l_1 regularization approach to model (1.4), the constrained convex optimization problem

$$\hat{\beta} = \underset{\beta}{\operatorname{argmin}}(\frac{1}{2n} \|y - Z\beta\|_2^2 + \lambda \|\beta\|_1), \text{ subject to } \sum_{j=1}^{p} \beta_j = 0, \tag{1.5}$$

is considered where $\beta = (\beta_1, ..., \beta_p)^T$, $\lambda > 0$ is a tuning parameter, and $\|.\|_2$ and $\|.\|_1$ denote the l_2 and l_1 norms, respectively. The zero-sum constraint in problem (1.5) is critical for the resulting estimator to have the following

desirable properties (Lin et al. (2014)).

1. *Scale invariance*: the estimator does not change under the transformation $X \mapsto TX$ for an arbitrary diagonal matrix $T = diag(t_1, ..., t_n)$ with all $t_i > 0$.

2. *Permutation invariance*: the estimator is invariant under any permutation π of the p components, meaning that it does not change if π is applied to both the columns of X and the components of $\hat{\beta}$.

3. *Selection invariance*: the estimator does not change if one knew ahead of time which components would be estimated as 0 and applied the procedure to the subcomposition formed by the components that remain.

Upon eliminating the constraint ($\sum_{j=1}^{p} \beta_j = 0$) by using $\beta_p = -\sum_{j=1}^{p-1} \beta_j$, problem (1.5) is rewritten as the unconstrained problem

$$\hat{\beta}_{-p} = \underset{\beta}{\operatorname{argmin}}(\frac{1}{2n} \left\| y - Z^p \beta_{-p} \right\|^2 + \lambda \left\| D\beta_{-p} \right\|_1),$$

where $\beta_{-p} = (\beta_1, ..., \beta_{p-1})^T$ and $D = (I_{p-1}, -1_p)^T \in \mathbb{R}^{p\text{x}(p-1)}$, with I_r, 1_r denotes the rxr identity matrix and the r-vector of ones, respectively (Lin et al. (2014)). However, Lin et al. (2014) found out that existing results did not specialize in their case to give an appropriate algorithm or theory for several reasons. First, getting rid of one arbitrary component and using a generic algorithm to the $(p-1)$-dimensional problem generally did not yield

numerical solutions that followed the permutation invariance property. Next, a version of the coordinate descent algorithm that is fast and applicable to a prespecified set of λ values was not yet available. Lastly, the generalized lasso problem theory did not provide useful insights into the compositional constraint and its effect on variable selection. The previous three limitations call for developing computational methods. One such method is the coordinate descent algorithm. These algorithms are known to be very efficient for solving large-scale l_1 regularization problems (Friedman et al. (2007)). However, they do not apply to problem (1.5) because the nondifferentiable l_1 terms are inseparable under the zero-sum constraint. Hence, Lin et al. (2014) proposed an efficient, easily implemented algorithm based on an iterative modification of coordinate descent that involves combining it with the method of multipliers or the augmented Lagrangian method (Bertsekas (2014)). Under their algorithm, the tuning parameter λ can be selected by the generalized information criterion (GIC) for high-dimensional penalized likelihood proposed by Fan and Tang (2013).

As mentioned before, Susin et al. (2020) focused on three methods for variable selection that acknowledged the compositional structure of microbiome data: selbal, clr-lasso, and coda-lasso. The method of selbal is a forward selection approach for the identification of compositional balances. Let $X = (X_1, ..., X_p)$ be the microbial composition of p taxa. Among these, we consider two disjoint subgroups of taxa, groups A and B, with p_A and p_B taxa indexed by $I_A \subset \{1, ..., p\}$ and $I_B \subset \{1, ..., p\}$, respectively, that do

not share taxa ($I_A \cap I_B = \emptyset$). The abundance balance between A and B, denoted by $\mathcal{B}(A, B)$, is defined as the log ratio between the geometric mean abundances of the two groups of taxa:

$$\mathcal{B}(A, B) = C * log(\frac{(\prod_{i \in I_A} X_i)^{1/p_A}}{(\prod_{j \in I_B} X_j)^{1/p_B}}),$$

where C is a normalization constant such that $C = \sqrt{\frac{p_A p_B}{p_A + p_B}}$. Selbal seeks for the two groups of taxa A and B whose relative abundances or balance $\mathcal{B}(A, B)$ is most associated with the outcome of interest Y according to the following generalized linear model:

$$g(E(Y)) = \beta_0 + \beta_1 \mathcal{B}(A, B) + \gamma^{'} Z,$$

where β_0 is the intercept, β_1 is the regression coefficient for the balance score, $Z = (Z_1, Z_2, ..., Z_r)$ are additional noncompositional covariates, γ is the vector of regression coefficients for Z and $\mathcal{B}(A, B)$ is defined as before. The optimal balance $\mathcal{B}(A, B)$ relies on identifying taxa that belong to either group A or B. The first step of this algorithm evaluates all possible taxa pairs to select the pair whose balance is most associated with the response. A forward selection process is performed where, at each step, a new taxon is added to the current balance, either in group A or B of the balance to improve the optimization criterion. The objective criterion is defined as the area under the receiver operating characteristic (ROC) curve, AUC, or the proportion

of explained deviance for a binary response and the mean squared error for a linear response. The algorithm stops when there is no remaining variable that improves the optimization criterion or when the maximum number of components in the balance, established with a cross-validation procedure, is reached.

Finally, clr-lasso and coda-lasso are two penalized regression models for compositional data analysis. For $(y_i, x_{1i}, ..., x_{pi}), i = 1, ..., n$, where y_i is the response and $x_i = (x_{1i}, ..., x_{pi})$ is the composition of p taxa for sample i, clr-lasso is defined as

$$y_i = \beta_0 + \beta_1 clr(x_{1i}) + ... + \beta_k clr(x_{pi}) + \epsilon_i. \tag{1.6}$$

The regression coefficients $\beta = (\beta_0, ..., \beta_k)$ are estimated to minimize

$$\sum_{i=1}^{n} (y_i - \beta_0 - \beta_1 clr(x_{1i}) - ... - \beta_k clr(x_{pi}))^2$$
$$\text{subject to } \sum_{j \geq 1} |\beta_j| < t \tag{1.7}$$

for a given constant t. Coda-lasso is formulated as

$$y_i = \beta_0 + \beta_1 log(x_{1i}) + ... + \beta_k log(x_{ki}) + \epsilon_i,$$

with constraint $\sum_{j \geq 1} \beta_j = 0$, where the regression coefficients $\beta = (\beta_0, ..., \beta_k)$

are estimated to minimize

$$\sum_{i=1}^{n}(y_i - \beta_0 - \beta_1 log(x_{1i}) - ... - \beta_k log(x_{ki}))^2$$
$$+\lambda \sum_{j\geq 1} |\beta_j| \text{ subject to } \sum_{j\geq 1} \beta_j = 0.$$

The minimization process is performed in two iterative steps based on soft thresholding and projection.

To summarize the results for Hron et al. (2013), Lin et al. (2014) and Susin et al. (2020), we have the following.

Omitting variables in compositional data analysis may lead to an enormous change in results from multivariate statistical analysis. In particular, this is the case for principal component analysis and the compositional biplot, where both the interpretation of loadings and scores of the remaining subcomposition are affected. The subcomposition is easier to handle and interpret. Hence, Hron et al. (2013) constructed a stepwise procedure that reduced the original composition to a subcomposition by avoiding a substantial change of the information, like those carried by the compositional biplot.

Motivated by research problems in analyzing gut microbiome and metagenomic data, Lin et al. (2014) considered variable selection and estimation in high-dimensional regression with compositional covariates, and proposed an l_1 regularization method for the linear log-contrast model that aligns with compositional data's unique features. Moreover, they formulated the pro-

posed procedure as constrained convex optimization and introduced a coordinate descent method of multipliers for efficient computation.

Though variable selection is one of the most relevant tasks in microbiome analysis, such as identifying microbial signatures, many studies still rely on methods that ignore the compositional nature of microbiome data. The applicability of compositional data analysis methods has been dampened by the availability of software and the difficulty in interpreting their results. Hence, Susin et al. (2020) came up with three methods for variable selection that acknowledge the compositional structure of microbiome data: selbal (a forward selection approach for the identification of compositional balances), and clr-lasso, coda-lasso (two penalized regression models for compositional data analysis).

Before moving into the main idea of this thesis, some advantages and disadvantages are highlighted here for each of the recently discussed works. For Hron et al. (2013), their stepwise procedure for excluding compositional parts allows for arriving at a subcomposition that retains the important information in the multivariate data structure. Their procedure aims to retain the total variance from one step to the next, which is stopped before a significant reduction occurs. In Lin et al. (2014), their model was shown to work in a plethora of scenarios (different (n,p and Σ combinations) and was illustrated to be useful in a microbiome study relating human body mass index to gut microbiome composition. On the contrary, they adopted their modeling approach in their microbiome data analysis because the total amount of

the microbiome data could not be reliably measured in experiments. If such measurements were available, it would be worth assuming a more flexible model where the total amount also plays a part in the response variable.

Lastly, in Susin et al. (2020), clr penalized regression is not subcompositionally consistent, meaning that different subcompositions will rise to different data transformations. Hence, results are not easily transferable from one study to another. Also, microbial signatures obtained from their approach can be challenging to implement on an independent dataset as it raises the question of how the variable from the new dataset should be clr-transformed and based on which components. It may be the case that the new dataset may include different components. On the contrary, penalized regression with coefficients restricted to a sum equal to zero, coda-lasso is an elegant and appropriate compositional data analysis approach. Computation time is efficient, and the results can be interpreted as balances between two groups of taxa with weights. Overall, for all three works, they had to use some transformation on X to get any meaningful result (Hron et al. (2013): clr, Lin et al. (2014): linear log, Susin et al. (2020): linear log), which motivated us to develop a transformation-free variable selection method for compositional data in this thesis.

For this thesis, we are particularly interested in a variable selection for compositional data in a linear regression setting. Given the compositional structure, removing any single compositional covariate, including those from the true underlying model, and performing linear regression on the remain-

ing covariates will not change the model fitting, including the coefficient of determination. Hence, to mitigate such a problem, we introduce a deletion done on pairs of covariates so the coefficient of determination is different from that of the original model. By implementing such a method, we can break the association among compositional covariates with the response variable and obtain more accurate results.

When the informative covariates have the same coefficients in the underlying true linear regression model, then the coefficient of determination of the original model will not differ from that of the model with the deleted pair of covariates, which skews the results. This is due to the unit-sum constraint placed on compositional covariates. Hence, we propose a deletion method on at least two covariates to help make our variable selection more efficient. For this method, we perform variable selection by focusing on two cases: 1) choosing a set of variables that show no difference in the coefficient of determination as compared to that of the original model, and 2) choosing a set of variables that show a decreased coefficient of determination as compared to that of the original model. We focused on the 1st case for convenience since it avoids over-selecting variables. We formulate this method as an algorithm to see how variable selection is carried out. We finish this thesis by carrying out simulation studies to show the effectiveness of our proposed algorithm in some instances and by illustrating the application of the proposed deletion method to a real data analysis.

Chapter 2

The proposed deletion method

In this chapter, we present the proposed deletion method to perform variable selection for compositional data. More specifically, first, we will delve into the important notation that will be used throughout the thesis. Next, we will detail how our proposed deletion method is constructed, starting with deleting two variables, etc. Finally, we develop an algorithm based on our theorem that will be used for simulation studies to evaluate the numerical performances of the developed method.

Suppose that we observe a $n \times 1$ vector y of responses and a $n \times p$ matrix $X = (X_1, X_2, ..., X_p)$ of covariates with each row of X being compositional and lying in $\mathbb{S}^{p-1}$, which is defined in chapter 1. Suppose that the model without any deleted variables is

$$y = \beta_0 + X\beta + \epsilon, \epsilon \sim N(0, \sigma^2) \tag{2.1}$$

where $\beta = (\beta_1, ..., \beta_p)'$ and ϵ is the random error term. Furthermore, let an informative covariate be defined as a covariate with a nonzero coefficient ($\beta_i \neq 0$ for $i = 1, ..., p$). From here, we are interested in performing variable selection to select the informative covariates for compositional data by seeing how the coefficient of determination behaves when we delete different variables in compositional data.

As a first step, we wanted to see how the coefficient of determination changes when we take away one compositional covariate at a time. Let X_1 represent the deleted covariate. Then, we have

$$\begin{aligned} y &= \beta_0 + \beta_1 X_1 + \beta_2 X_2 + ... + \beta_p X_p + \epsilon \qquad (2.2) \\ &= \beta_0 + \beta_1 (1 - \sum_{i=2}^{p} X_i) + \beta_2 X_2 + ... + \beta_p X_p + \epsilon \\ &= \beta_0 + \beta_1 - \beta_1 \sum_{i=2}^{p} X_i + \beta_2 X_2 + ... + \beta_p X_p + \epsilon \\ &= \beta_0 + \beta_1 + (\beta_2 - \beta_1) X_2 + (\beta_3 - \beta_1) X_3 + ... + (\beta_p - \beta_1) X_p + \epsilon. \qquad (2.3) \end{aligned}$$

If we let $\beta_0^* = \beta_0 + \beta_1, \beta_2^* = \beta_2 - \beta_1, \beta_3^* = \beta_3 - \beta_1, ..., \beta_p^* = \beta_p - \beta_1$, then we have that

$$y = \beta_0^* + \beta_2^* X_2 + \beta_3^* X_3 + ... + \beta_p^* X_p + \epsilon \qquad (2.4)$$

Hence, the model fitting of (2.4) does not change from the model fitting of (2.2). Thus, this leads to the R^2 remaining the same. If we were to

delete X_2, X_3,..., X_p individually, the same thing would happen. Thus, when deleting one covariate at a time from the original model, the coefficient of determination remains the same. Hence, we propose to start with deleting a pair of covariates instead. More details on this proposed approach will be explained in the next section.

2.1 Proposed method

Remember that an informative covariate is a covariate that is part of the true model with a nonzero regression coefficient, while a non-informative covariate has zero coefficients in the model. Again, when we focus on deleting one covariate at a time, we see that the coefficient of determination remained the same, as shown in the previous subsection.

As explained above, we could perform this variable selection by either (1) choosing a group of covariates that deleting them in pairs results in a difference in R^2 close to 0 or (2) choosing a group of covariates that deleting them in pairs results in a difference in R^2 that is far from 0. For this thesis, we went with the first option since we know both variables in option (1) are noninformative, and we only know one in option (2) is informative, but we cannot know which one is. Keep in mind that if the original linear regression model contains no intercept, we should have a deletion method on one covariate since we do not have the collinearity problem regardless of the $\sum_{i=1}^{p} x_i = 1$ constraint. In contrast, if the original data matrix $X_{n\times p}$ contains

an intercept, we perform the deletion method on at least two covariates given the reason derived in Equations (2.2)-(2.4).

2.2 Implementation algorithm for $X_{n\times p}$

Given the covariate matrix $X_{n\times p}$ and the response vector $y_{n\times 1}$, the developed algorithm proceeds as follows.

- *Step* 0: remove any column in X, and run linear regression of y on the remaining $p-1$ columns of X with intercept. The resulting coefficient of determination is denoted as R^2_O.
- *Step* 1: Let Ω_1 be an empty set.
- *Step* 2: Let $k = 2$.
- *Step* 3: For each subset of k columns of X, $X_i^k = (X_{i_1}, \ldots X_{i_k})$ for $i = 1, 2, \ldots, C_k^p$, delete X_i^k in X, and denote the resulting covariate matrix as X_i^*. Regress y on X^* with intercept, and denote the resulting coefficient of determination as $R^2_{p_i}$.
- *Step* 4: Compute the difference $d_i^k = R^2_{p_i} - R^2_O$, for $i = 1, 2, \ldots, C_2^p$.
- *Step* 5: Select the non-informative variable set $\Omega_k = \{X_i^k : |d_i^k| \leq \delta \text{ for } i = 1, 2. \ldots, C_k^p\}$, where δ is a small threshold value.
- *Step* 6: If $\Omega_k = \Omega_{k-1}$, go to Step 8.

- *Step* 7: Let $k = k + 1$ and repeat Steps 2-5 till $k = p$.

- *Step* 8: Stop.

The final selected non-informative covariate set is Ω_k, and the selected informative set is the complement set Ω_k^c. In the algorithm above, we increase the size of the deleted subsets to accommodate the cases where some nonzero coefficients have the same value. In the underlying linear model (**??**). If the coefficients of the informative variables are all different, the developed algorithm above should stop at $k = 3$ and $\Omega_2 = \Omega_3$ is the set of selected non-informative variables. When only two nonzero coefficients are the same, the algorithm will stop at $k = 4$ and $\Omega_3 = \Omega_4$ is the set of selected non-informative variables. The trend continues with more nonzero coefficients having the same value.

Keep in mind that if we were to introduce another algorithm but focusing on $X_{n\times p}$ not having an intercept column, then all of the steps would remain the same, except step 3 would involve deleting one covariate instead since under solving for X_i by the unit-sum constraint, X_i (regular ith compositional covariate) and X_i^* (deleted ith compositional covariate) would not be the same.

More details on how the optimal δ is chosen are explained in the following section.

2.3 Choosing the optimal δ criterion

Two measures of model selection accuracy that we use are FP (number of false positives) and FN (number of false negatives). In our variable selection setting, an FP corresponds to a noninformative variable identified as an informative variable, and an FN corresponds to an informative variable identified as a noninformative variable. We want to see how these accuracy measures relate to the tuning parameter δ, which, as shown later, is chosen by the BICc.

Keep in mind that δ is used by way of $|d_i| \leq \delta$ to identify the pairs of noninformative compositional variables, where d_i denotes the R^2-difference between the models with and without the corresponding subset of compositional variables. This selection tool says that if the associated absolute value of the R^2-difference is less than δ, then the ith deleted subset of variables is selected as noninformative variables. This signifies that if we have a large enough δ, we choose mainly the informative variables. More specifically, as the δ increases, all the noninformative variables will be identified correctly, which, in turn, causes the FP to decrease. However, if we have a small enough δ, we choose all the variables as informative variables. More specifically, as the δ decreases, all variables will be selected as informative variables, which, in turn, causes the FN to decrease.

The optimal δ is chosen using the corrected Bayesian Information Criterion (BICc) to balance FN and FP. The idea of it is as follows. We first carry

out the variable selection on all possible δ values as a sequence of quantiles of the R^2 differences from all possible deletions. We do this so that there are 100 different δ values (1st δ corresponds to 1%, 2nd δ corresponds to 2%, etc.). Then, from there, for each δ value, we implement the above algorithm (with intercept column) and calculate the BICc for each model considering the final chosen important variables. The BICc is defined as follows (McQuarrie (1999)).

$$BICc = log(\hat{\sigma}_p^2) + log(n)\frac{p}{n-p-2} \tag{2.5}$$

where n is the sample size, p is the number of coefficients (including intercept) and $\hat{\sigma}_p^2 = \frac{RSS}{n}$ such that $RSS = \sum_{i=1}^{n}(y_i - \hat{y}_i)^2$ is the residual sum of squares. We choose the δ value with the lowest BICc value.

Recall that the BIC is defined as

$$BIC = log(\hat{\sigma}_p^2) + log(n)\frac{p}{n} \tag{2.6}$$

where n, p and $\hat{\sigma}_p^2$ are defined as before. For the BIC, $log(\hat{\sigma}_p^2)$ decreases much faster than $log(n)p/n$ increases leading to a global minimum at the saturated model, which in turn leads to a higher probability of overfitting (McQuarrie (1999)). On the contrary, for the BICc, $log(\hat{\sigma}_p^2)$ decreases much slower than $log(n)p/(n-p-2)$ increases, leading to a general balance at the saturated model which in turn leads to a lower probability of overfitting (McQuarrie (1999)). Moreover, in small samples, the BIC can overfit more than

the BICc (McQuarrie (1999)) due to their signal-to-noise ratios. Remember that a weak signal-to-noise ratio for overfitting indicates a higher probability of overfitting. Refer to McQuarrie (1999) for more details.

2.4 Proposed method for $n < p$ case

Our developed variable selection procedure for compositional data is presented above, assuming $n > p$. For high dimensional compositional data with $n < p$ or ultra-high dimensional data with $n << p$, we used the Sure Independence Screening (SIS) method (Fan and Lv (2008)) to reduce the data dimensionality first before carrying out our proposed method. The idea behind this method is the following. We initially center and scale the columns $x_1, x_2, ..., x_p$ from X such that the mean is 0 and sample standard deviation is 1 (Fan and Lv (2008)). Let $M_* = \{1 \leq i \leq p : \beta_i \neq 0\}$ be the true sparse model with nonsparsity rate $s = |M_*|$ (Fan and Lv (2008)). The other $p - s$ variables can be correlated with y by linkage to the predictors in the model. Let $\omega = (\omega_1, \omega_2, ..., \omega_p)^T$ be a p-vector obtained by $\omega = X^T y$, where the $n \times p$ data matrix X is first standardized by column. For any given $\gamma \in (0, 1)$, the p componentwise magnitudes of the vector ω are sorted in decreasing order and we define a submodel $M_\gamma = \{1 \leq i \leq p : |\omega_i|$ is among the first $[\gamma n]$ largest of all$\}$, where $[\gamma n]$ denotes the integer part of γn.

By the end, we shrink the model $\{1, 2, ..., p\}$ down to a submodel M_γ with

size $d = [\gamma n] < n$ (Fan and Lv (2008)) where $n > d$. Its computational cost is multiplying a $p \times n$ matrix with an n-vector plus getting the largest d components of a p-vector, which signifies that SIS has computational complexity $O(np)$. We specify that d is below the sample size n. We let $d = k * n/log(n)$. To ensure that all informative variables were initially selected, we consider $k = 1$, $k = 2$, and $k = 3$.

SIS has the property that all the informative variables survive after variable selection, with probability tending to one, which, in turn, narrows down the search for informative predictors. Furthermore, SIS can reduce dimensionality from high up to an exponential growth (e.g., $exp(O(n^{\xi}))$ for some $\xi > 0$) to a relatively large scale d (e.g., $o(n)$) that is below sample size. It can not only speed up variable selection rapidly but also improve the estimation accuracy when the dimensionality is ultra-high. SIS combined with well-developed lower dimensional techniques such as the SCAD, Lasso, or adaptive Lasso provides a powerful tool for high dimensional variable selection (Fan and Lv (2008)). As shown later in this thesis, it proves to be quite helpful in our simulation studies.

Chapter 3

Imputed KNN algorithm

3.1 Missing Data

Most statistical methods cannot be directly applied to data sets with missing observations. Even though the observations with missing information could be deleted in the univariate case, this can result in a severe loss of information in the multivariate case (Hron et al. (2010)). Multivariate observations usually form the rows of a data matrix, and deleting an entire row implies that cells carrying available information are lost for the analysis. In both the univariate and multivariate case, the problem remains that valid inferences can only be made if the missing data are *missing completely at random* (MCAR) (Little et al. (2019)). Instead of deleting observations with missing values, filling in the missing cells with appropriate values is also considered. For the multivariate case, this is only possible provided additional informa-

tion is available. Once all missing values have been imputed, the data set can be analyzed using the standard techniques for complete data.

Many different methods for imputation have been developed over the last few decades. While univariate methods replace the missing values with the coordinate-wise mean or median, the more appropriate methods are based on similarities among the objects and/or variables (Hron et al. (2010)). A typical distance-based method is k-nearest neighbor (KNN) imputation, where the information of the nearest $k \geq 1$ complete observations is used to estimate the missing values. Another well-known procedure is the expectation maximization (EM) algorithm (Dempster et al. (1977)), which uses the relations between observations and variables to estimate the missing cells in a data matrix. These methods can deal with both MCAR and *missing at random* (MAR) missing values mechanisms (Little et al. (2019)). Furthermore, one usually assumes that the data originates from a multivariate normal distribution, which is no longer valid when outliers are present in the data. Hence, the classical methods can give very biased estimates for the missing values, and it is more advisable to use robust methods, being less influenced by outlying observations (Béguin et al. (2008); Serneels et al. (2008)). It turned out that classical or robust imputation methods worked well for standard multivariate data, i.e. for data with a direct representation in the Euclidean space (Yucel et al. (2010)). Unfortunately, this is not the case with compositional data. Hence, we need a method that can handle such data.

In the following paragraphs, we revisit and apply the KNN imputation

method to compositional data, focusing on single and multiple imputation cases. Section 3.2 focuses on the single imputation case. In contrast, Section 3.3 focuses on the multiple imputation case, as later on, we evaluated the improvements of the multiple imputation upon the single imputation.

3.2 Single Imputation

We implemented an imputed k nearest neighbors (KNN) algorithm for our second project on our full dataset. In the past, KNN imputation was proved to be successful for standard multivariate data (Troyanskaya et al. (2001)). We adapt the KNN technique on multivariate compositional data with missing observations. The idea is to find the k most similar observations to a missing composition by an appropriate distance measure and then replace these missing compositions with the available variable information of the identified nearest neighbors. We chose Aitchison's distance (Aitchison (1986)) as the distance measure in this scenario. For two compositions $\mathbf{x} = (x_1, ..., x_p)^t$ and $\mathbf{y} = (y_1, ..., y_p)^t$, the Aitchison's distance is defined as

$$d_A(\mathbf{x}, \mathbf{y}) = \sqrt{\frac{1}{p} \sum_{i=1}^{p-1} \sum_{j=i+1}^{p} (ln(\frac{x_i}{x_j}) - ln(\frac{y_i}{y_j}))^2}. \tag{3.1}$$

Replacing the Euclidean distance with the Aitchison distance is crucial because the simplex space has a different geometrical structure than the classical Euclidean space. Principles on this geometry were introduced in Aitchi-

son (1986), and the resulting so-called Aitchison geometry holds the vector space as well as Hilbert space properties (Egozcue and Pawlowsky-Glahn (2006)). This allows the construction of a basis on the simplex, and consequently, standard statistical methods designed for the Euclidean space can be applied. One such basis that came to mind was the isometric logratio (ilr) transformation (Egozcue, Pawlowsky-Glahn et al. (2003)). The ilr transformation results in a $p-1$ dimensional real space, offering good theoretical and practical properties (Egozcue and Pawlowsky-Glahn (2005)). One important property is the isometry, meaning the Aitchison distance of two compositions $\mathbf{x}$ and $\mathbf{y}$ is the same as the ordinary Euclidean distance d_E for their ilr images $ilr(\mathbf{x})$ and $ilr(\mathbf{y})$, i.e.

$$d_A(\mathbf{x}, \mathbf{y}) = d_E(ilr(\mathbf{x}, \mathbf{y})). \tag{3.2}$$

Thus, the ilr transformation allows to represent compositional data in terms of the standard Euclidean geometry, and therefore standard statistical methods can be applied.

For our KNN algorithm, the imputation can be done sequentially (one observation after the other) within a composition with several missing observations by searching the k-nearest neighbors among observations where all information corresponding to the non-missing observations and that in the variable to be imputed is available. We use this approach instead of imputing simultaneously for all observations because, in general, the k observations

can change during sequential imputation as compared with simultaneous imputation, as this approach uses the information of the same k observations. Furthermore, more neighbors will be considered for imputation, and requesting more information per observation will lead to more reliable imputation results. Our KNN algorithm is explained in more detail in the following paragraph.

Let us consider a composition $x_i = (x_{i1}, ..., x_{ip})^t, i = 1, ..., n$, with n being the number of observations and p be the total number of variables in x_i. Then, More specifically, our KNN imputation algorithm goes as follows.

1. Let $O = C^c \cup R^c$ such that $C^c = \{1, ..., p\} \backslash C$ and $R^c = \{1, ..., n\} \backslash R$ where $C \subset \{1, ..., p\}$ and $R \subset \{1, ..., n\}$ denote the column and row, respectively, indices of the missing cells for x_i

2. For any $i \in R$ and $j \in C$, consider among all the remaining compositions those which have non-missing parts at positions i, j and O, and compute the k-nearest neighbors $x_{i_1}, ..., x_{i_k}$ to the composition x_i using the Aitchison distance.

3. The jth cell of all k-nearest neighbors is of interest for imputation. Hence, the imputed value replacing the missing cell x_{ij} is $x^*_{ij} = median(x_{i_1 j}, ..., x_{i_k j})$.

4. To make sure the unit-sum constraint is satisfied, we set $x^{**}_{ij} = (1 - \sum_{o \in O} x_{io}) \frac{x^*_{ij}}{\sum_{i \in R, j \in C} x^*_{ij}}$ where x^{**}_{ij} represents our final imputed value.

As mentioned above, for step 1, we use the distance measure of the Aitchison distance instead of the Euclidean distance because the simplex sample space has a different geometrical structure than the classical Euclidean space. Furthermore, again, note that the Aitchison distance of two compositions is the same as the ordinary Euclidean distance for their ilr images. Next, in step 3, we used the median instead of the mean of the k-nearest neighbors as the median mitigates the effect of outliers compared with the mean. Moreover, we do not have enough information to justify that our data comes from a symmetric distribution. With the imputed values for missing data using this algorithm, we then proceed with our proposed deletion method.

3.3 Multiple Imputation

Unfortunately, there are problems with using single imputation. Using this method often results in an underestimation of the variability because each unobserved value carries the same weight in the analysis as the known, observed values Jakobsen et al. (2017). Furthermore, the validity of this method does not depend on whether data are MCAR but instead on specific assumptions that the missing values, for example, are identical to the last observed value Jakobsen et al. (2017). These assumptions are often unrealistic, and hence, single imputation is often a potentially biased method and should be used with great caution. Hence, we aim to improve the results by using multiple imputation to our algorithm from before.

The idea of the multiple imputation algorithm is as follows. Assume the same setup as before. Then:

1. For l=1,2,..,N
   (a) sample the non-missing parts of X by row and call it X_l where X_l is the lth sampled non missing version of X.
   (b) For X_l, run the single imputation as before where we are imputing a missing cell $x_{ij(l)}$ instead.
   (c) We impute the missing cell x_{ij} by the median (i.e. $x^*_{ij(l)} = median(x_{i_1 j(l)}, ..., x_{i_k j(l)})$).
2. By the end, we impute x_{ij} by the mean (i.e. $x^{**}_{ij} = mean(x^*_{ij(1)}, ..., x^*_{ij(N)})$).
3. Set $x^{***}_{ij} = (1 - \sum_{o \in O} x_{io}) \frac{x^{**}_{ij}}{\sum_{i \in R, j \in C} x^{**}_{ij}}$ to make sure the unit-sum constraint is satisfied where x^{***}_{ij} represents our final imputed value.

For step 1 (a), we sampled only the non-missing parts of X while keeping the positions of the missing parts of X the same, as we wanted to make sure that the positions of the missing parts remained the same for each imputation so the single imputation algorithm can stay consistent for each imputation. In step (2), we took the mean of the medians of the k-nearest neighbors for all missing observations in terms of all imputations, as we wanted to make sure that our final imputed value depended on a range of values instead of on the values of its closest non-missing observations.

Multiple imputation is a feasible, credible, and powerful approach to handling missing data that helps reduce bias in several scenarios (Enders (2017)).

Multiple imputation attempts to minimize the impact of nonresponse bias on the analysis by using available information about individuals to adjust the parameter estimates (Woods et al. (2021)). Using multiple imputation thus approximates what results would look like with complete observations while allowing for representation of uncertainty in the results and maximizing the dataset's statistical power (Cheema (2014); Dong et al. (2013))

Chapter 4

Simulation Studies

We conducted simulation studies on our developed deletion method to evaluate its numerical performance on variable selection and compare it with other methods for compositional data under various settings. The methods that we focused on were Lin et al. (2014) and Susin et al. (2020).

We generated the covariate variables in the following way. We first generated a $n \times p$ data matrix $W = (w_{ij})$ by generating n random observations from a multivariate normal distribution $N_p(\underline{0}, \Sigma)$, and then obtained the covariate matrix $X = (x_{ij})$ by the transformation $x_{ij} = exp(w_{ij}) / \sum_{k=1}^{p} exp(w_{ik})$ where $\underline{0}$ is the $p \times 1$ zero vector, and $\Sigma = (\rho^{|i-j|})$ with $\rho = 0.2, 0.5$ or 0.9. We generated the responses according to model (**??**) where $\beta_0 = 3$ and $\beta^* = (1, -0.8, 0.6, 0, 0, -1.5, -0.5, 1.2, 0, ..., 0)^T$. Additionally, for the imputed data, we set $k = 30$, the percentage of missing rows to $(5\%, 10\%, 15\%, 20\%)$, and the number of missing covariates per observation to $(4, 5, 6, 7)$.

We set $(n, p) = (50, 30), (100, 200)$ and $(100, 1000)$, and repeated 100 simulations for each setting. When $n < p$, we proceeded with the SIS method using $d = j*(n/log(n))$ as described in section 3.3 where $j = 1, 2$ or 3, before implementing the developed method. The tuning parameter δ in the developed method was selected by the corrected Bayesian Information Criterion (BICc) as described in section 3.2. We used two performance measures for our comparisons: the total number of false negatives (FN) and false positives (FP), where positives and negatives refer to nonzero and zero coefficients, respectively. The random errors were generated from the normal distribution $N(0, \sigma^2)$ where σ^2 was specified to control the varied signal noise ratio (SNR) (Johnson (2006)), which is defined as

$$SNR = \frac{Var(X\beta)}{\sigma^2}.$$

where σ^2 is the variance of ϵ. The means and standard errors of these performance measures for this model are in Tables (4.1)-(4.4).

Table 4.1: means and standard errors (in parentheses) of FP/FN for the developed deletion, Lin et al. (2014) and Susin et al. (2020) methods based on 100 full datasets for $(n, p) = (50, 30)$ where the mean and SE of the SNR are, respectively, 15.41 and 0.67

method	technique	FP	FN
delet.	NA	1.32 (0.29)	0.44 (0.09)
Lin	NA	3.80 (0.29)	0.00 (0.00)
Susin	selbal	2.37 (0.36)	1.71 (0.16)
Susin	clr-lasso	6.37 (0.50)	0.85 (0.14)
Susin	coda-lasso	8.19 (0.60)	0.73 (0.14)

Table 4.2: means and standard errors (in parentheses) of FP/FN for the developed deletion method based on 100 full datasets for $(n, p) = (100, 200)$, $d = (21, 42, 63)$ where the mean and SE of the SNR are, respectively, 18.32 and 0.84

d	FP	FN	FP (SIS)	FN (SIS)
21	0.78 (0.12)	0.85 (0.08)	15.85 (0.08)	0.85 (0.08)
42	1.20 (0.15)	0.48 (0.06)	36.47 (0.06)	0.47 (0.06)
63	1.10 (0.20)	0.40 (0.06)	57.36 (0.06)	0.36 (0.06)

Table 4.3: means and standard errors (in parentheses) of FP/FN for the developed deletion and Lin et al. (2014) methods based on 100 full datasets for $(n, p) = (100, (200, 1000))$, $d = 63$, $\rho = (0.2, 0.5, 0.9)$ where the mean and SE of the SNR are, respectively, 18, and between 0.70 and 1.33

method	n	p	ρ	FP	FN	FP (SIS)	FN (SIS)
delet.	100	200	0.2	1.39 (0.19)	0.54 (0.07)	57.48 (0.06)	0.48 (0.06)
Lin	100	200	0.2	3.17 (0.20)	0.00 (0.00)	NA (NA)	NA (NA)
delet.	100	200	0.5	1.47 (0.19)	1.23 (0.11)	57.94 (0.08)	0.94 (0.08)
Lin	100	200	0.5	5.88 (0.26)	0.00 (0.00)	NA (NA)	NA (NA)
delet.	100	200	0.9	1.38 (0.19)	2.04 (0.15)	58.28 (0.10)	1.28 (0.10)
Lin	100	200	0.9	17.20 (0.53)	0.00 (0.00)	NA (NA)	NA (NA)
delet.	100	1000	0.2	3.25 (0.40)	1.53 (0.10)	58.38 (0.09)	1.38 (0.09)
Lin	100	1000	0.2	1.52 (0.19)	0.00 (0.00)	NA (NA)	NA (NA)
delet.	100	1000	0.5	4.51 (0.47)	2.64 (0.10)	59.06 (0.08)	2.06 (0.08)
Lin	100	1000	0.5	7.40 (0.45)	0.00 (0.00)	NA (NA)	NA (NA)
delet.	100	1000	0.9	4.10 (0.39)	2.97 (0.16)	59.06 (0.13)	2.06 (0.13)
Lin	100	1000	0.9	3.26 (0.43)	4.60 (0.14)	NA (NA)	NA (NA)

Table 4.4: means and standard errors (in parentheses) of FP/FN for my deletion and Lin et al. (2014) methods based on 100 imputed datasets for $(n, p) = (100, 200)$, $d = 63$ where the SNR ranges between 18.20 and 18.32

method	RM%	CM	FP	FN	FP (SIS)	FN (SIS)
delet.	5	4	1.05 (0.19)	0.39 (0.06)	57.35 (0.06)	0.35 (0.06)
delet.	5	5	1.08 (0.19)	0.40 (0.06)	57.36 (0.06)	0.36 (0.06)
delet.	5	6	1.22 (0.23)	0.39 (0.06)	57.35 (0.06)	0.35 (0.06)
delet.	5	7	1.15 (0.21)	0.40 (0.07)	57.36 (0.06)	0.36 (0.06)
delet.	10	4	1.13 (0.18)	0.39 (0.06)	57.35 (0.06)	0.35 (0.06)
delet.	10	5	1.29 (0.21)	0.40 (0.06)	57.37 (0.06)	0.37 (0.06)
delet.	10	6	1.26 (0.19)	0.41 (0.06)	57.36 (0.06)	0.36 (0.06)
delet.	10	7	1.14 (0.21)	0.40 (0.06)	57.36 (0.06)	0.36 (0.06)
delet.	15	4	1.02 (0.17)	0.39 (0.06)	57.36 (0.06)	0.36 (0.06)
delet.	15	5	1.26 (0.23)	0.40 (0.06)	57.35 (0.06)	0.35 (0.06)
delet.	15	6	1.04 (0.18)	0.37 (0.06)	57.34 (0.06)	0.34 (0.06)
delet.	15	7	1.11 (0.20)	0.45 (0.07)	57.39 (0.06)	0.39 (0.06)
delet.	20	4	1.35 (0.21)	0.38 (0.07)	57.35 (0.06)	0.35 (0.06)
delet.	20	5	1.20 (0.19)	0.39 (0.06)	57.36 (0.06)	0.36 (0.06)
delet.	20	6	1.06 (0.17)	0.37 (0.06)	57.34 (0.06)	0.34 (0.06)
delet.	20	7	1.07 (0.16)	0.37 (0.06)	57.33 (0.06)	0.33 (0.06)
Lin	5	4	2.78 (0.23)	0.00 (0.00)	NA (NA)	NA (NA)
Lin	5	5	2.80 (0.24)	0.00 (0.00)	NA (NA)	NA (NA)
Lin	5	6	2.67 (0.22)	0.00 (0.00)	NA (NA)	NA (NA)
Lin	5	7	2.72 (0.23)	0.00 (0.00)	NA (NA)	NA (NA)
Lin	10	4	2.69 (0.21)	0.00 (0.00)	NA (NA)	NA (NA)
Lin	10	5	2.76 (0.24)	0.00 (0.00)	NA (NA)	NA (NA)
Lin	10	6	2.82 (0.23)	0.00 (0.00)	NA (NA)	NA (NA)
Lin	10	7	2.75 (0.22)	0.00 (0.00)	NA (NA)	NA (NA)
Lin	15	4	2.49 (0.22)	0.00 (0.00)	NA (NA)	NA (NA)
Lin	15	5	2.59 (0.24)	0.00 (0.00)	NA (NA)	NA (NA)
Lin	15	6	2.70 (0.22)	0.00 (0.00)	NA (NA)	NA (NA)
Lin	15	7	2.80 (0.22)	0.00 (0.00)	NA (NA)	NA (NA)
Lin	20	4	2.74 (0.21)	0.00 (0.00)	NA (NA)	NA (NA)
Lin	20	5	2.79 (0.23)	0.00 (0.00)	NA (NA)	NA (NA)
Lin	20	6	2.71 (0.22)	0.00 (0.00)	NA (NA)	NA (NA)
Lin	20	7	2.58 (0.22)	0.00 (0.00)	NA (NA)	NA (NA)

RM% = % of rows missing
CM= # of columns missing

4.1 Lin et al. (2014) vs. Susin et al. (2020) vs. deletion method

First, we compared our proposed deletion method with both Lin et al. (2014) and Susin et al. (2020) methods under the $(n, p) = (50, 30)$ setting. The results for this comparison are summarized in Table (4.1). For this comparison, we set $\sigma^2 = 6.4 \times 10^{-4}$, leading to the mean and standard error of the SNR to be 15.41 and 0.67, respectively. Our proposed deletion and Lin's methods performed much better than Susin's method as our deletion method's and Lin's FN are much smaller than their FN. Hence, we decided not to present further simulation results with Susin's method to save space. Compared with the method of Lin et al. (2014), our method had a lower FP. However, since we cared more about the FN due to the presence of important variables, Lin's method performed better than our proposed deletion method due to the FN for Lin's method being 0 and our FN being close to 0.45. Hence, we performed more simulations on Lin's method to get a bigger idea of how well or worse Lin's method performs against our deletion method.

4.2 Comparing our deletion method when varying d from SIS

Next, we focused on our deletion method using the full dataset by varying the parameter d in SIS to see how the FN changes and later choosing the optimal

d under the $(n, p) = (100, 200)$ setting. The results for this comparison are in Table (4.2). For this comparison, we set $\sigma^2 = 4.4 \times 10^{-5}$, resulting in the SNR's mean and standard error being 18.32 and 0.84, respectively. Note that since the FN after performing both SIS and our deletion method is similar to the FN after performing SIS, the final FN is mainly caused by SIS, not by our developed deletion method. Initially, we focused on $d = 21$. However, seeing that the FN was close to 1, we gradually increased to $d = 42$ and $d = 63$. The case of $d = 63$ proved to be good enough as the FN was close to 0. By the end, we noticed that the FN was lower as the d increased, meaning that our deletion method performed better. Since $d = 63$ had the best FN out of all d values we tried, we later chose $d = 63$ for further simulation comparisons under the $n < p$ case.

4.3 Lin et al. (2014) vs. deletion methods when varying ρ

For our third comparison, we varied the value of ρ in Σ to see when our deletion method performed better than Lin's method using the complete dataset for both the $(n, p) = (100, 200)$ and $(100, 1000)$ cases. The results for this comparison are in Table (4.3). For $(n, p) = (100, 200)$, we set σ^2 to be 1.24×10^{-5}, 1.06×10^{-5} and 3.07×10^{-5} representing ρ being 0.2, 0.5 and 0.9, respectively. For $(n, p) = (100, 1000)$, we set σ^2 to be 5.01×10^{-7}, 4.38×10^{-7} and 1.25×10^{-7} corresponding to ρ being 0.2, 0.5 and 0.9, respectively. For

both cases, these σ^2 values lead to the mean and standard error of the SNR to be 18 and [0.70,1.33], respectively. For the $(n, p) = (100, 200)$ case, as the ρ increased, our method's FP remained between 1.40 and 1.50 while Lin's FP got progressively worse, whereas our FN got worse while Lin's FN remained at 0. Hence, our deletion method performed better than Lin's regarding the FP. Furthermore, for the $(n, p) = (100, 1000)$ case, as the ρ increased, our method's and Lin's FN increased to the point where our method's FN was lower than Lin's FN, indicating our deletion method performed better than Lin's method in this setting. Overall, Lin's method performed worse for both cases when ρ got near 1. Moreover, as the ρ increased, our deletion method performed better than Lin's.

4.4 Full vs. imputed datasets for deletion method

We compared our full and imputed datasets against each other using our deletion method. Looking at Tables (4.2) and (4.4) under $d = 63$, we see that the FN for the imputed dataset regardless of what setting is used is slightly lower than that for the full dataset. Hence, our imputed dataset performed better than our full dataset under our deletion method.

4.5 Lin et al. (2014) vs. deletion methods for imputed dataset

Lastly, we compared our deletion method against Lin's using the imputed dataset and $d = 63$ for the $(n, p) = (100, 200)$ setting. The results for this comparison are in Table (4.4). We set $\sigma^2 = 4.4 \times 10^{-5}$, resulting in the mean and standard error of the SNR to be [18.20,18.32] and 0.84, respectively. Looking at the first half of Table (4.4), our deletion method performed exceptionally well since it ended with an FN close to 0 regardless of the setting. Even though Lin's method performed better than our deletion method in terms of the FN, our deletion method had an FN, again, close to 0 and performed better than Lin's method for the FP.

Looking at all comparisons, overall, our deletion method performed better than Susin's method, whereas, under specific settings, our deletion method performed better than Lin's method. Note that the FN for Lin's method always was 0 unless we go to an extreme setting as demonstrated in Table (4.3). However, our deletion method, in general, performed better on FP than Lin's method.

4.6 Single vs. Multiple KNN Imputation

To refine our deletion method when applied to data with missing observations, we also show how well our deletion method performed with both single

and multiple KNN imputations to see if, by the end, multiple imputation improved the results further. We used the same setting as before, where $(n, p) = (50, 30), (100, 200)$. Also, we used 1, 3, 5, and 7 imputations where 1 corresponds to single imputation and 3, 5, and 7 correspond to multiple imputations. We focused on 15% rows and four missing observations for each subject for our data and other missing settings yield similar results, and thus were omitted. Moreover, for $(n, p) = (100, 200)$, we used $d = 63$ since this d value performed the best according to Table (4.2). For our comparisons, we used the same performance measures as before. Table (4.5) summarizes the means and standard errors of these models' performance measures.

Table 4.5: means and standard errors (in parentheses) of FP/FN for single vs. multiple kNN imputation (15% rows and 4 columns missing data) for our deletion method where the SNR ranges between 15.13 and 15.21
Imp = #of imputations

n	p	d	Imp	FP	FN	FP (SIS)	FN (SIS)
50	30	NA	1	3.61 (0.69)	0.31 (0.08)	NA	NA
50	30	NA	3	3.35 (0.61)	0.28 (0.07)	NA	NA
50	30	NA	5	3.56 (0.58)	0.40 (0.08)	NA	NA
50	30	NA	7	3.21 (0.61)	0.32 (0.08)	NA	NA
100	200	63	1	2.84 (0.46)	0.37 (0.06)	57.35 (0.06)	0.35 (0.06)
100	200	63	3	2.54 (0.39)	0.40 (0.06)	57.36 (0.06)	0.36 (0.06)
100	200	63	5	2.15 (0.35)	0.38 (0.06)	57.35 (0.06)	0.35 (0.06)
100	200	63	7	2.24 (0.33)	0.37 (0.06)	57.34 (0.06)	0.34 (0.06)

Looking at Table (4.5), for $(n, p) = (50, 30)$ and $(100, 200)$, as the number of imputations increased, the FP and FN remained roughly constant. However, for both cases, by the end, the FP substantially decreased when comparing one and seven imputations. Hence, for the FP, multiple impu-

tation helped a little bit. However, for the FN, it did not, and thus, the developed procedure is relatively robust against the number of imputations. If we were to increase the number of imputations beyond seven imputations, these results could further be improved. However, due to computational cost, we could not increase the number of imputations further.

Chapter 5

Data Analysis

Exclusive breast-feeding is the preferred method of feeding during the first 6 months of age to support optimal growth and development and to protect against gastrointestinal disease, diarrhea, and respiratory tract infection. It is the reference model against which all alternative feeding methods are measured with regard to growth, health, development, and all other short-term and long-term outcomes (Gartner et al. (2005)). In recent years, extensive research has been geared towards the lipid component of breast milk, which provides not only calories and macronutrition but also key micronutrients for infant growth and cognitive development. Nayak et al. (2017) considered the lipid composition of breast milk's impact on early infant growth and cognitive development particularly from low-income populations in the Indian subcontinent.

A single breast milk specimen was collected within six weeks postpar-

tum from two low-income maternal cohorts of exclusively breastfed infants from Dhaka, Bangladesh (n =683) and Kolkata, India (n =372), and assayed for percentage composition of 26 FAs (fatty acids). Individual FAs were expressed as percentage wt/(wt of total identified FA); the FA profile of each specimen contained 26 individual FAs. Other variables considered for this study included maternal age, maternal anthropometry (maternal height/weight), maternal BMI, maternal education, expenditure, infant birth order, days of lactation, total number of pregnancies, and infant age at breast milk collection. For our project, we considered maternal height, maternal education, expenditure, and infant birth order since those variables significantly affected the response variable as in Nayak et al. (2017). The response variables for this project were the waz, haz, and whz, which stand for the weight for age, height for age, and weight for height, respectively, assessed using the Multicentre Growth Reference Study (MGRS) application. Lastly, we focused on the following periods: 24 weeks, 52 weeks, and 104 weeks.

Using the previously mentioned information, we compared our proposed deletion, Lin's and Susin's methods, by seeing which final FAs out of the 26 FAs affected the waz, haz, and whz using the model of waz, haz or whz vs. the 26 FAs, maternal height, maternal education, expenditure, and infant birth order within the previously mentioned periods where Table (5.1) has each of the 26 FAs' notation. Note that the extra variables are kept regardless of which final FAs were selected. We focused on choosing the optimal δ through

the AIC, BIC, and BICc for our deletion method. Furthermore, we wanted to see which method(s) performed the best by comparing their associated MSEs from the models using the selected variables.

Table 5.1: The 26 fatty acids and their descriptions

FA	Description
CAP	Capric
LAU	Lauric
MYR	Myristic
PAL	Palmitic
PLA	Palmitelaidic
PLE	Palmitoleic
STE	Stearic
ELA	Elaidic
OLE	Oleic
LLA	Linoelaidic
LA	Linoleic
ARA	Arachidic
GLA	gamma-Linolenic
EIC9	Eicosenoic
ALA	alpha-Linolenic
EDA	Eicosadienoic
BEH	Behenic
DGLA	Dihomo-g-linolenic
AA	Arachidonic
LIG	Lignoceric
EPA	Eicosapentaenoic
NER	Nervonic
DTA	Docosatetraenoic
DPA6	Docosapentaenoic-n6
DPA	Docosapentaenoic-n3
DHA	Docosahexaenoic

For our deletion method, we initially conducted an F-test for all the y/period combinations between the models of y vs. X and y vs. Z_{-1}, X where Z_{-1} represents the remaining FAs after a specific one of them is taken out. For any such y/period combinations with a p-value less than 0.05, we proceed with the data analysis for the deletion method. Otherwise, for such y/period combinations with a p-value of at least 0.05, we end with the y vs. X model and calculate its associated MSE. Table (5.2) summarizes the results for this initial step. Looking at Table (5.2), for the waz during 52 and 104 weeks and haz during 24, 52, and 104 weeks, since they have p-values less than 0.05, we proceeded with the data analysis for our deletion method. For the whz during 24, 52, and 104 weeks, we ended with the y vs. X model for our deletion method. Hence, by the end, for our deletion method, we see a relationship between the 26 FAs and the waz and haz.

As mentioned above, we wanted to compare our proposed deletion with Lin's and Susin's methods by seeing what final FAs out of the 26 FAs are selected for waz, haz, and whz. For our deletion method, we focused on the waz during 52 and 104 weeks and haz during 24, 52, and 104 weeks since their associated F-test p-values were less than 0.05. Remember that if any method ends up with no FAs, we choose the model of y vs. X. For our deletion method, the boxplots of the R^2-differences and optimal δ selected by the AIC, BIC, BICc are shown in Figure (5.1) a)-e). For the boxplots, remember that if the optimal δ is within the R^2-differences, then at least one of the FAs will be chosen at the end. If the optimal δ is the maximum

of all the R^2-differences, then all FAs will be treated as noninformative by the end, whereas if the optimal δ is the minimum of all the R^2-differences, then all FAs will be treated as informative by the end. The results for this comparison are summarized in Tables (5.3)-(5.11).

For our deletion method, 5 out of 6 cases ended with the BIC/BICc not choosing any FAs, whereas for the haz during 104 weeks case, the BIC/BICc only chose the DPA FA. Moreover, the AIC chose at least two FAs for all the pertinent cases. Looking at Table (5.4), all methods chose the EIC9. All methods picked the PAL, ELA, OLE, and NER FAs based on Table (5.5). Proceeding with Table (5.6), all methods chose the ARA and BEH FAs. All methods picked the ARA and NER FAs when looking at Table (5.7). Finally, glancing at Table (5.8), all methods picked the GLA, EDA, and DPA FAs. Hence, for all methods, the Eicosenoic, Palmitic, Elaidic, Oleic, and Nervonic fatty acids have an effect on the weight for age using the MGRS application when the maternal height, maternal education, expenditure, and infant birth order are kept. Meanwhile, for all methods, the Arachidic, Behenic, Nervonic, gamma-linolenic, Eicosadienoic, and Docosapentaenoic-n3 fatty acids affect the height for age using MGRS application when the maternal height, maternal education, expenditure, and infant birth order are kept.

Looking at Tables (5.4)-(5.8) again, we see that each method had almost roughly the same MSEs where Susin's method had the lowest MSEs of them all. We notice this same trend in Tables (5.3) and (5.9)-(5.11) as well. How-

ever, remember that it had the lowest MSE because it chose the most FAs by the end.

As a last step to this data analysis, we wanted to explore other studies that used this dataset to verify if some of the FAs that were chosen before had an impact still. One noticeable study is Yakes et al. (2011). This study examined the cross-sectional relationship between prolonged breastfeeding and maternal BMI, assessed the adequacy of fat intake among lactating and non-lactating mothers of children 24-48 months of age, and determined breast milk FA composition. Dietary data were collected during two non-consecutive 24-hour periods via 12-hour in-home daytime observations and recall. The National Cancer Institute method for episodically consumed foods was used to estimate usual intake distributions. By the end of the study, almost all women were estimated to consume less than the recommended intake levels for total LA, GLA, AA, DPA, DGLA, DHA, and ALA. Hence, the Linoleic, gamma-linolenic, Arachidonic, Docosapentaenoic-n3, Dihomo-g-linolenic, Docosahexaenoic, and alpha-linolenic fatty acids affect this study. Compared to our data analysis, for all three methods, the Arachidonic, gamma-linolenic, and Docosapentaenoic-n3 fatty acids were found to have an effect as well.

Another noticeable study is Szabó et al. (2010). This study compared the fatty acid composition of human milk at two different stages of lactation. It investigated the relationship between trans isomeric and long-chain polyunsaturated fatty acids (LCPUFAs) in human milk at the sixth month

of lactation. Human milk samples were obtained from 462 mothers who participated in a large birth cohort study at the sixth week and sixth month of lactation. The fatty acid composition of human milk lipids was determined by high-resolution capillary gas-liquid. It was shown that the percentage contributions to human milk fatty acid composition of PUFAs (LA, AA, ALA, and DHA) increased significantly. Hence, the Linoleic, Arachidonic, alpha-linolenic, and docosahexaenoic fatty acids affect this study. Compared to our data analysis, the Arachidonic fatty acids were also found to affect all three methods.

We conclude that Arachidonic is one of the most essential fatty acids for a woman at birth to digest due to it having an effect in all the previous studies, including ours. This fatty acid is involved in early neurological development as infants.

y	week	p-value
waz	24	0.0570
waz	52	0.0234
waz	104	0.0105
haz	24	0.0028
haz	52	0.0013
haz	104	0.0079
whz	24	0.3589
whz	52	0.2173
whz	104	0.1004

Table 5.2: p-values of F-test between models of y vs. Z,X and y vs. X where y represents the waz,haz or whz, Z represents the 26 FAs and X represents htcm,medu,bodr,exp

	method					
	deletion			Susin		
	AIC	BIC/BICc	Lin	selbal	clr-lasso	coda-lasso
CAP					x	
LAU						
MYR					x	
PAL						
PLA						
PLE			x	x	x	x
STE			x	x	x	x
ELA						
OLE				x	x	
LLA						
LA						
ARA						
GLA			x	x	x	x
EIC9			x	x	x	x
ALA						
EDA				x	x	
BEH				x	x	
DGLA						
AA				x	x	
LIG						
EPA						
NER						x
DTA						
DPA6						
DPA					x	x
DHA				x	x	
MSE	1.2761	1.2761	1.2418	1.1924	1.1844	1.2367

x = chosen FA

Table 5.3: selected FAs, and MSEs from final linear regression model of either y vs. X or y vs. Z, X for the waz variable during 24 weeks for the deletion (AIC,BIC,BICc for optimal δ), Lin's and Susin's methods

	method					
	deletion			Susin		
	AIC	BIC/BICc	Lin	selbal	clr-lasso	coda-lasso
CAP						
LAU						
MYR						
PAL			x	x		
PLA						
PLE			x	x	x	
STE						
ELA						
OLE				x		
LLA	x					
LA					x	
ARA			x		x	x
GLA			x	x	x	x
EIC9	x		x	x	x	
ALA	x					
EDA				x		
BEH				x		x
DGLA						
AA				x		
LIG			x	x	x	x
EPA	x			x		
NER			x	x	x	x
DTA				x		
DPA6						
DPA			x	x	x	x
DHA				x	x	
MSE	0.4819	0.5263	0.4812	0.4372	0.4768	0.4913

x = chosen FA

Table 5.4: selected FAs, and MSEs from final linear regression model of either y vs. X or y vs. Z, X for the waz variable during 52 weeks for the deletion (AIC,BIC,BICc for optimal δ), Lin's and Susin's methods

	method					
	deletion			Susin		
	AIC	BIC/BICc	Lin	selbal	clr-lasso	coda-lasso
CAP	x					x
LAU						
MYR	x					x
PAL	x		x			x
PLA						
PLE						x
STE	x					
ELA	x		x			x
OLE	x		x		x	x
LLA	x					
LA			x		x	
ARA				x		
GLA			x		x	x
EIC9			x	x	x	x
ALA	x					
EDA				x	x	x
BEH						
DGLA				x	x	x
AA	x					x
LIG						
EPA	x					x
NER	x		x		x	x
DTA	x			x		x
DPA6						
DPA			x	x	x	x
DHA					x	x
MSE	0.9396	1.0415	0.9576	1.0002	0.9467	0.9191

x = chosen FA

Table 5.5: selected FAs, and MSEs from final linear regression model of either y vs. X or y vs. Z, X for the waz variable during 104 weeks for the deletion (AIC,BIC,BICc for optimal δ), Lin's and Susin's methods

	method					
	deletion			Susin		
	AIC	BIC/BICc	Lin	selbal	clr-lasso	coda-lasso
CAP					x	x
LAU						
MYR			x	x	x	x
PAL				x	x	x
PLA						
PLE					x	x
STE			x	x	x	x
ELA					x	x
OLE						
LLA				x	x	x
LA				x	x	x
ARA	x			x	x	x
GLA			x	x	x	x
EIC9				x	x	x
ALA					x	x
EDA				x	x	x
BEH	x			x	x	x
DGLA			x			x
AA				x	x	x
LIG						
EPA					x	x
NER			x		x	x
DTA					x	x
DPA6					x	x
DPA					x	x
DHA						
MSE	0.8476	0.8490	0.8382	0.8049	0.7505	0.7505

x = chosen FA

Table 5.6: selected FAs, and MSEs from final linear regression model of either y vs. X or y vs. Z, X for the haz variable during 24 weeks for the deletion (AIC,BIC,BICc for optimal δ), Lin's and Susin's methods

	method					
	deletion			Susin		
	AIC	BIC/BICc	Lin	selbal	clr-lasso	coda-lasso
CAP						
LAU						
MYR			x		x	
PAL						
PLA						x
PLE					x	
STE						x
ELA						
OLE						
LLA	x		x			
LA						
ARA	x		x	x	x	x
GLA			x	x	x	x
EIC9						
ALA	x					
EDA			x		x	x
BEH						
DGLA	x					
AA						x
LIG	x				x	x
EPA	x					
NER	x		x		x	x
DTA						
DPA6						
DPA					x	x
DHA			x	x	x	x
MSE	0.1586	0.1597	0.1568	0.1590	0.1554	0.1557

x = chosen FA

Table 5.7: selected FAs, and MSEs from final linear regression model of either y vs. X or y vs. Z, X for the haz variable during 52 weeks for the deletion (AIC,BIC,BICc for optimal δ), Lin's and Susin's methods

	method					
	deletion			Susin		
	AIC	BIC/BICc	Lin	selbal	clr-lasso	coda-lasso
CAP						
LAU						
MYR			x	x	x	x
PAL			x			x
PLA						
PLE						
STE						x
ELA						
OLE				x		
LLA				x		x
LA						
ARA						
GLA	x		x	x	x	x
EIC9						x
ALA						
EDA	x		x		x	x
BEH	x					
DGLA			x		x	x
AA				x	x	x
LIG						x
EPA				x		
NER			x		x	x
DTA						
DPA6	x					
DPA	x	x	x	x	x	x
DHA						
MSE	0.8169	0.8312	0.7744	0.7819	0.7728	0.7590

x = chosen FA

Table 5.8: selected FAs, and MSEs from final linear regression model of either y vs. X or y vs. Z, X for the haz variable during 104 weeks for the deletion (AIC,BIC,BICc for optimal δ), Lin's and Susin's methods

	method					
	deletion			Susin		
	AIC	BIC/BICc	Lin	selbal	clr-lasso	coda-lasso
CAP					x	
LAU					x	
MYR						
PAL					x	
PLA					x	
PLE			x	x	x	
STE						
ELA						
OLE					x	
LLA						
LA					x	
ARA						
GLA			x	x	x	
EIC9			x	x	x	
ALA						
EDA					x	
BEH				x	x	
DGLA					x	
AA						
LIG					x	
EPA			x			
NER				x	x	
DTA					x	
DPA6				x	x	
DPA						
DHA				x	x	
MSE	1.1243	1.1243	1.0767	1.0594	1.0049	1.1243

x = chosen FA

Table 5.9: selected FAs, and MSEs from final linear regression model of either y vs. X or y vs. Z, X for the whz variable during 24 weeks for the deletion (AIC,BIC,BICc for optimal δ), Lin's and Susin's methods

	method					
	deletion			Susin		
	AIC	BIC/BICc	Lin	selbal	clr-lasso	coda-lasso
CAP						
LAU						
MYR						
PAL				x		
PLA						
PLE			x	x	x	x
STE						
ELA					x	
OLE				x		
LLA						
LA				x		
ARA				x		
GLA			x		x	x
EIC9			x	x	x	x
ALA						
EDA						
BEH				x	x	x
DGLA				x		
AA				x		
LIG					x	
EPA			x			
NER				x	x	x
DTA				x		x
DPA6				x	x	
DPA			x	x	x	x
DHA				x		
MSE	0.8337	0.8337	0.7470	0.6823	0.7247	0.7310

x = chosen FA

Table 5.10: selected FAs, and MSEs from final linear regression model of either y vs. X or y vs. Z, X for the whz variable during 52 weeks for the deletion (AIC,BIC,BICc for optimal δ), Lin's and Susin's methods

	method					
	deletion			Susin		
	AIC	BIC/BICc	Lin	selbal	clr-lasso	coda-lasso
CAP				x	x	
LAU						
MYR						
PAL				x		
PLA						
PLE			x	x	x	
STE						
ELA						
OLE			x	x	x	
LLA						
LA			x	x	x	
ARA						
GLA				x		
EIC9			x	x	x	x
ALA						
EDA						
BEH				x		
DGLA				x		
AA				x		
LIG				x		
EPA						
NER				x		
DTA				x		
DPA6				x	x	
DPA					x	x
DHA			x	x	x	x
MSE	0.9149	0.9149	0.8407	0.7957	0.8384	0.8795

x = chosen FA

Table 5.11: selected FAs, and MSEs from final linear regression model of either y vs. X or y vs. Z, X for the whz variable during 104 weeks for the deletion (AIC, BIC, BICc for optimal δ), Lin's and Susin's methods

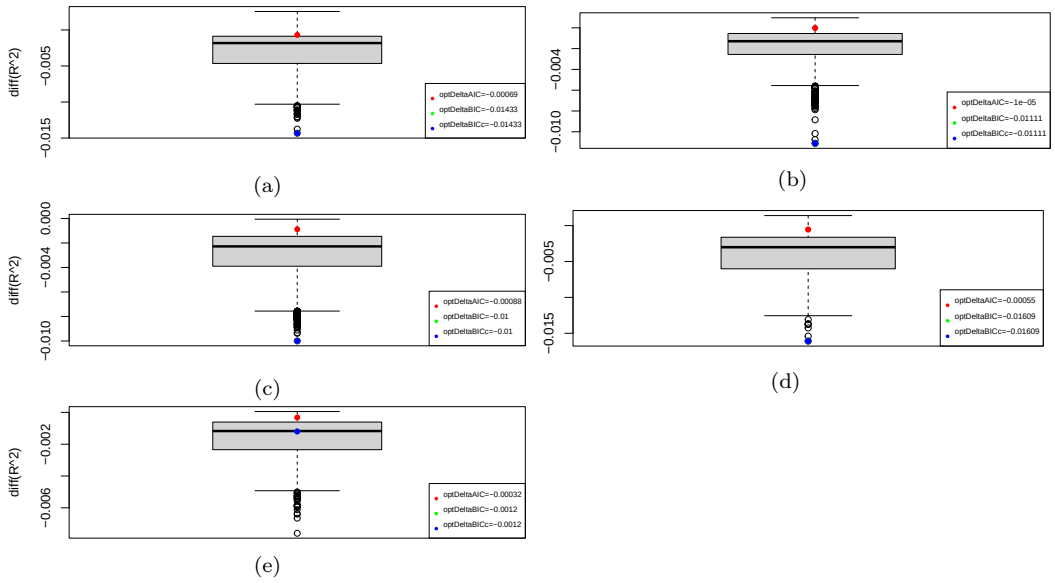

Figure 5.1: boxplots of R^2-differences for the models of waz vs. Z, X during a) 52 weeks, b) 104 weeks, and haz vs. Z, X during c) 24 weeks, d) 52 weeks, e) 104 weeks

Chapter 6

Future Work

Some of the work we plan to do for this project in the future includes using parallel computing for our deletion method, extending it to generalized linear model settings for our deletion method, and implementing a clustering technique for both our single and multiple imputation methods.

- We plan to implement parallel computing to our deletion method to address the involved combinatorics problem when p is large. More specifically, we plan to use parallel computing to delete a pair, triple,..., of variables simultaneously, as it would help speed up the computation of our deletion method since the computation of the R^2 difference of deleting a set of variables does not depend on the deletion of another set of variables. The way we would do this is the following. First, we would use the `doParallel` and `foreach` packages in R. Under the `doParallel` package, we use the `registerDoParallel` func-

tion to set the number of cores used to the maximum allowed as that would allow our deletion method to run on multiple cores instead of one core. Then, under the `foreach` package, we would replace all of the pertinent `for` loops with the `foreach` loop, `%dopar%` and `%:%` combination to take care of deleting a pair, triple, etc. of covariates.

The first to second-to-last `foreach` loops with have the `%:%` at the end as the `%:%` operator is known as a nesting operator that turns multiple `foreach` loops into a single loop. This combination will accumulate all of the indices of the variables that will be deleted. The last `foreach` loop will have the `%dopar%` at the end as the `%dopar%` makes use of the `registerDoParallel` function to run the pertinent task(s) under multiple cores. This combination will take care of deleting the pertinent variables from X. All of this pertinent code will be set equal to a new variable X_d where, from there, we carry the deletion method as we would typically do.

- We want to extend our deletion method to generalized linear models (GLM) instead. to perform variable selection for compositional data in more complicated settings. Consider the model without any deleted variables to be

$$g(E(y)) = \beta_0 + X\beta,\ y \sim F$$

where β_0, X and β are defined as before, $g(\cdot)$ is a link function that

specifies how $E(y)$ relates to X, and F is the distribution of y in the exponential distribution family. For example, if $y \sim Poisson$, we can choose the link function $g(E(y)) = log(E(y))$ such that the model is

$$ln(E(y)) = \beta_0 + X\beta, \tag{6.1}$$

whereas if $y \sim Binomial$, we can choose the link function $g(E(y)) = logit(E(y))$ such that the model is

$$logit(E(y)) = \beta_0 + X\beta.$$

Since we are working with GLMs, we will use the difference in deviances between the original model and the model after deleting compositional variables to find which variables are informative. The deviance of a GLM is defined as

$$D = 2 \times log(\frac{\mathcal{L}(\hat{\theta}_1, ..., \hat{\theta}_N)}{\mathcal{L}(\hat{\beta})})$$

where $log(\mathcal{L}(\hat{\theta}_1, ..., \hat{\theta}_N))$ is the maximum value of the log-likelihood function for the saturated model such that $\theta_i \propto g^{-1}(\beta_0 + X\beta)$, for all $i = 1, ..., N$, and $log(\mathcal{L}(\hat{\beta}))$ is value of the maximized log-likelihood function when fitting model (6.1). We will evaluate different criteria, such as the AIC, BIC, etc., on choosing the optimal δ to see which gives us an optimal δ that leads to the best results.

- We want to revisit K-means clustering and add it to our single KNN imputation. Consider the same setup as that for the single KNN imputation. Then:

  1. Run steps 1 and 2 from the single KNN imputation
  2. Divide the k-nearest neighbors of x_i into K clusters
  3. For $c = 1, ..., K$:
     (a) We are interested in imputing $x_{ij(c)}$. Hence, replace $x_{ij(c)}$ by $x^*_{ij(c)} = median(x_{i_1 j(c)}, ..., x_{i_k j(c)})$.
     (b) To make sure the unit-sum constraint is satisfied, we set $x^{**}_{ij(c)} = (1 - \sum_{o \in O} x_{io(c)}) \frac{x^*_{ij(c)}}{\sum_{i \in R, j \in C} x^*_{ij(c)}}$.
  4. Use an appropriate criterion to determine which c out of the K clusters had the best-imputed value.
  5. For the cth value that had the best-imputed value, our final imputed value becomes $x^{**}_{ij(c)}$ from step 3(b).

  Note that we might end up with potential outliers when we stick with the original k-nearest neighbors for a missing value. Hence, to make sure that there are no outliers, we wanted to divide these k-nearest neighbors into K roughly equal groups. Note that $K \leq p$. For a specific row of x_i, if one observation is missing, $K = p - 1$ whereas if more than one observation is missing, we will investigate how to specify K. Lastly, when we are done with the single imputation case, we also

want to try K-means clustering on multiple imputation as well to see if the multiple imputation results improve.

- We want to revisit and redefine our multiple imputation algorithm for improved numerical results in Table (4.5).

  Assuming the same setup for the single imputation algorithm where $n > p$, we may proceed as follows.

  1. For $l = 1, 2, ..., N$

     (a) Run step 1 (a)-(c) from the previous multiple imputation algorithm where X_l^* denotes the lth sampled non-missing version of X after we impute $x_{ij(l)}$ accordingly.

     (b) Remove any column in X_l^* and run linear regression of y on the remaining $p-1$ columns of X_l^* with intercept. The resulting coefficient of determination is denoted as R_O^2.

     (c) Let $\Omega_{1(l)}$ be an empty set and $k_l = 2$.

     (d) For each subset of k_l columns of X_l^*, $X_{i(l)}^{k_l} = (X_{i_1(l)}^*, ..., X_{i_{k_l}(l)}^*)$ for $i = 1, 2, ..., C_{k_l}^p$, delete $X_{i(l)}^{k_l}$ in X_l^*, and denote the resulting covariate matrix as $X_{i(l)}^{**}$. Regress y on X_l^{**} with intercept, and denote the resulting coefficient of determination as $R_{p_i}^2$.

     (e) Compute $d_i^{k_l}$ the same way that d_i^k was calculated from our deletion method.

     (f) Select the non-informative variable set $\Omega_{k_l(l)} = \{X_{i(l)}^{k_l} : |d_i^{k_l}| \leq$

δ for $i = 1, 2, ..., C_{k_l}^{p}\}$, where δ is defined and chosen the same way as that from our deletion method.

(g) If $\Omega_{k_l(l)} = \Omega_{(k_l-1)(l)}$, go to Step (i).

(h) Let $k_l = k_1 + 1$ and repeat Steps (c)-(f) till $k_l = p$.

(i) Stop.

2. Use the majority vote among $\Omega_{k_1(1)}, \Omega_{k_2(2)}, ..., \Omega_{k_N(N)}$ to see which final variables appear in at least half of these sets where such a final set is denoted as Ω_{all}.

Note that if $n < p$, the algorithm is the same as before, but for each imputation, we perform 1 (a) first, followed by SIS, and then 1 (b)-(i) instead.

- For the $n < p$ case, we consider using a more efficient ultra-high dimensional data screening technique than SIS to improve our deletion method's FN in the simulation studies. Note that, in SIS, when all covariates are standardized, ranking them in order of (absolute) correlation with the response is equivalent to ordering the estimated slopes $|\hat{\beta}_j|$ (Ghosh et al. (2021)). However, SIS is non-robust since the estimates $\hat{\beta}_j'$s are from MLE/OLS. Ghosh et al. (2021) used the same approach as SIS, but with robust estimates for β_j in the marginal model using the density power divergence (DPD) approach. Let us fix a $j \in \{1, 2, ..., p\}$

and $\alpha > 0$. Consider the jth marginal model

$$y_i = \gamma_j + \beta_j x_{ij} + \epsilon_j, i = 1, ..., n \tag{6.2}$$

where $\epsilon \sim N(0, \sigma_j^2)$, X_j is the jth covariate for each $j = 1, ..., p$, and $\theta_j = (\gamma_j, \beta_j, \sigma_j)^T$ are estimated by the MLE/OLS methods. Based on the marginal model (6.2), the objective function can be simplified to the form $H_{n,\alpha}(\theta_j) = \frac{1}{n}\sum_{i=1}^{n} l_\alpha(y_i, \gamma_j + \beta_j x_{ij}, \sigma_j)$, where

$$l_\alpha(y, \eta, \sigma) = \frac{1}{\sigma^\alpha (2\pi)^{\alpha/2}}\left(\frac{1}{\sqrt{1+\alpha}} - \frac{1+\alpha}{\alpha} e^{\frac{-\alpha(y-\eta)^2}{\sigma^2}}\right) + \frac{1}{\alpha}.$$

Then, θ_j is estimated as $\hat{\theta}_j^M$ such that

$$\hat{\theta}_j^M = (\hat{\gamma}_j^{M\alpha}, \hat{\beta}_j^{M\alpha}, \hat{\sigma}_j^{M\alpha}) = \underset{\theta}{\operatorname{argmin}} \frac{1}{n}\sum_{i=1}^{n} l_\alpha(y_i, \gamma_j + \beta_j x_{ij}, \sigma).$$

Based on $\hat{\theta}_j^M$, for a given $\alpha > 0$, we can choose the informative variables in order of the values of $|\hat{\beta}_j^{M\alpha}|$, which is referred to as the DPD-SIS procedure (Ghosh et al. (2021)). Note that at $\alpha = 0$ (in a limiting sense), $\hat{\theta}_j^M$ coincides with $\hat{\theta}_j$. Thus, the DPD-SIS algorithm at $\alpha = 0$ becomes identical to the SIS. The extent of robustness of the DPD-SIS increases with increasing $\alpha > 0$.

References

Aitchison, J and Sheng M Shen (1980). "Logistic-normal distributions: Some properties and uses". In: *Biometrika* 67(2), pg. 261–272.

Aitchison, John (1982). "The statistical analysis of compositional data". In: *Journal of the Royal Statistical Society: Series B (Methodological)* 44(2), pg. 139–160.

Aitchison, John (1986). *The Statistical Analysis of Compositional Data*. Chapman and Hall.

Aitchison, John and John Bacon-Shone (1984). "Log contrast models for experiments with mixtures". In: *Biometrika* 71(2), pg. 323–330.

Aitchison, John and Michael Greenacre (2002). "Biplots of compositional data". In: *Journal of the Royal Statistical Society Series C: Applied Statistics* 51(4), pg. 375–392.

Béguin, Cédric and Beat Hulliger (2008). "The BACON-EEM algorithm for multivariate outlier detection in incomplete survey data". In: *Survey Methodology* 34(1), pg. 91.

Bertsekas, Dimitri P (2014). *Constrained optimization and Lagrange multiplier methods*. Academic press.

Bühlmann, Peter and Sara Van De Geer (2011). *Statistics for high-dimensional data: methods, theory and applications*. Springer Science & Business Media.

Cheema, Jehanzeb R (2014). "Some general guidelines for choosing missing data handling methods in educational research". In: *Journal of Modern Applied Statistical Methods* 13(2), pg. 3.

Dempster, Arthur P, Nan M Laird and Donald B Rubin (1977). "Maximum likelihood from incomplete data via the EM algorithm". In: *Journal of the royal statistical society: series B (methodological)* 39(1), pg. 1–22.

Dong, Yiran and Chao-Ying Joanne Peng (2013). "Principled missing data methods for researchers". In: *SpringerPlus* 2, pg. 1–17.

Egozcue, Juan José and Vera Pawlowsky-Glahn (2005). "Groups of parts and their balances in compositional data analysis". In: *Mathematical Geology* 37(7), pg. 795–828.

Egozcue, Juan José, Vera Pawlowsky-Glahn, Glòria Mateu-Figueras and Carles Barcelo-Vidal (2003). "Isometric logratio transformations for compositional data analysis". In: *Mathematical geology* 35(3), pg. 279–300.

Egozcue, Juansnm José and Vera Pawlowsky-Glahn (2006). "Simplicial geometry for compositional data". In: *Geological Society, London, Special Publications* 264(1), pg. 145–159.

Enders, Craig K (2017). "Multiple imputation as a flexible tool for missing data handling in clinical research". In: *Behaviour research and therapy* 98, pg. 4–18.

Fan, Jianqing and Jinchi Lv (2008). "Sure independence screening for ultrahigh dimensional feature space". In: *Journal of the Royal Statistical Society: Series B (Statistical Methodology)* 70(5), pg. 849–911.

Fan, Yingying and Cheng Yong Tang (2013). "Tuning parameter selection in high dimensional penalized likelihood". In: *Journal of the Royal Statistical Society: Series B (Statistical Methodology)* 75(3), pg. 531–552.

Filzmoser, Peter (1999). "Robust principal component and factor analysis in the geostatistical treatment of environmental data". In: *Environmetrics: The Official Journal of the International Environmetrics Society* 10(4), pg. 363–375.

Filzmoser, Peter, Karel Hron and Clemens Reimann (2009). "Principal component analysis for compositional data with outliers". In: *Environmetrics: The Official Journal of the International Environmetrics Society* 20(6), pg. 621–632.

Fišerová, Eva and Karel Hron (2011). "On the interpretation of orthonormal coordinates for compositional data". In: *Mathematical Geosciences* 43(4), pg. 455.

Friedman, Jerome, Trevor Hastie, Holger Höfling, Robert Tibshirani et al. (2007). "Pathwise coordinate optimization". In: *The annals of applied statistics* 1(2), pg. 302–332.

Gartner, Lawrence M et al. (2005). "Breastfeeding and the use of human milk." In: *Pediatrics* 115(2), pg. 496–506.

Ghosh, Abhik and Magne Thoresen (2021). "A robust variable screening procedure for ultra-high dimensional data". In: *Statistical Methods in Medical Research* 30(8), pg. 1816–1832.

Greenacre, Michael (2019). "Variable selection in compositional data analysis using pairwise logratios". In: *Mathematical Geosciences* 51(5), pg. 649–682.

Hron, Karel et al. (2010). "Imputation of missing values for compositional data using classical and robust methods". In: *Computational Statistics & Data Analysis* 54(12), pg. 3095–3107.

Hron, Karel et al. (2013). "Covariance-based variable selection for compositional data". In: *Mathematical geosciences* 45(4), pg. 487–498.

Jakobsen, Janus Christian, Christian Gluud, Jørn Wetterslev and Per Winkel (2017). "When and how should multiple imputation be used for handling missing data in randomised clinical trials–a practical guide with flowcharts". In: *BMC medical research methodology* 17(1), pg. 1–10.

Johnson, Don H (2006). "Signal-to-noise ratio". In: *Scholarpedia* 1(12), pg. 2088.

Lin, Wei, Pixu Shi, Rui Feng and Hongzhe Li (2014). "Variable selection in regression with compositional covariates". In: *Biometrika* 101(4), pg. 785–797.

Little, Roderick JA and Donald B Rubin (2019). *Statistical analysis with missing data.* **volume** 793. John Wiley & Sons.

Maronna, Ricardo A, R Douglas Martin, Victor J Yohai and Matıas Salibián-Barrera (2019). *Robust statistics: theory and methods (with R)*. John Wiley & Sons.

McQuarrie, Allan D (1999). "A small-sample correction for the Schwarz SIC model selection criterion". In: *Statistics & probability letters* 44(1), pg. 79–86.

Nayak, Uma et al. (2017). "Influence of maternal and socioeconomic factors on breast milk fatty acid composition in urban, low-income families". In: *Maternal & child nutrition* 13(4), e12423.

Real, Carlos, J Ángel Fernández, Jesús R Aboal and Alejo Carballeira (2011). "Substituting missing data in compositional analysis". In: *Environmental pollution* 159(10), pg. 2797–2800.

Serneels, Sven and Tim Verdonck (2008). "Principal component analysis for data containing outliers and missing elements". In: *Computational Statistics & Data Analysis* 52(3), pg. 1712–1727.

Susin, Antoni, Yiwen Wang, Kim-Anh Lê Cao and M Luz Calle (2020). "Variable selection in microbiome compositional data analysis". In: *NAR Genomics and Bioinformatics* 2(2), lqaa029.

Szabó, Éva et al. (2010). "Fatty acid profile comparisons in human milk sampled from the same mothers at the sixth week and the sixth month of lactation". In: *Journal of pediatric gastroenterology and nutrition* 50(3), pg. 316–320.

Troyanskaya, Olga et al. (2001). "Missing value estimation methods for DNA microarrays". In: *Bioinformatics* 17(6), pg. 520–525.

Wang, Huiwen, Liying Shangguan, Junjie Wu and Rong Guan (2013). "Multiple linear regression modeling for compositional data". In: *Neurocomputing* 122, pg. 490–500.

Woods, Adrienne D et al. (2021). "Best practices for addressing missing data through multiple imputation". In: *Infant and child development*, e2407.

Yakes, Elizabeth A et al. (2011). "Intakes and breast-milk concentrations of essential fatty acids are low among Bangladeshi women with 24–48-month-old children". In: *British Journal of Nutrition* 105(11), pg. 1660–1670.

Yucel, Recai M and Hakan Demirtas (2010). "Impact of non-normal random effects on inference by multiple imputation: A simulation assessment". In: *Computational statistics & data analysis* 54(3), pg. 790–801.