Noname manuscript No.
(will be inserted by the editor)

Robust regression with compositional covariates including cellwise outliers

Nikola Štefelová · Andreas Alfons · Javier Palarea-Albaladejo · Peter Filzmoser · Karel Hron

Received: date / Accepted: date

Abstract We propose a robust procedure to estimate a linear regression model with compositional and real-valued explanatory variables. The proposed procedure is designed to be robust against individual outlying cells in the data matrix (cellwise outliers), as well as entire outlying observations (rowwise outliers). Cellwise outliers are first filtered and then imputed by robust estimates. Afterwards, rowwise robust compositional regression is performed to obtain model coefficient estimates. Simulations show that the procedure generally outperforms a traditional rowwise-only robust regression method (MM-estimator). Moreover, our procedure yields better or comparable results to recently proposed cellwise robust regression methods (shooting S-estimator, 3-step regression) while it is preferable for interpretation through the use of appropriate coordinate systems for compositional data. An application to bio-environmental data reveals that the proposed procedure—compared to other regression methods—leads to conclusions that are best aligned with established scientific knowledge.

Keywords Cellwise outliers · Compositional data · Logratio coordinates · Regression analysis · Robust statistics

Subject classification 62J05, 62H99.

N. Štefelová
Palacký University, Faculty of Science, 17. listopadu 12, 77146 Olomouc, Czech Republic
E-mail: nikola.stefelova@seznam.cz

A. Alfons
Erasmus Universiteit Rotterdam, Econometric Institute, PO Box 1738, 3000 DR Rotterdam, The Netherlands
E-mail: alfons@ese.eur.nl

J. Palarea-Albaladejo
Biomathematics and Statistics Scotland, JCMB, The King's Buildings, Peter Guthrie Tait Road, Edinburgh, EH9 3FD, Scotland, UK
E-mail: javier.palarea@bioss.ac.uk

P. Filzmoser
Vienna University of Technology, Institute of Statistics and Mathematical Methods in Economics, Wiedner Hauptstraße 8-10, A-1040 Vienna, Austria
E-mail: peter.filzmoser@tuwien.ac.at

K. Hron
Palacký University, Faculty of Science, 17. listopadu 12, 77146 Olomouc, Czech Republic
E-mail: hronk@seznam.cz

1 Introduction

Regression analysis is one of the most widely used techniques in practical data analysis and statistical modelling. It allows to study how a real-valued response variable is associated with explanatory variables of various types, including variables of a compositional nature (i.e., variables that carry relative information). Compositional variables are commonly generated through some form of signal processing in modern areas of chemistry, biology and environmental sciences. They are expressed in units such as percentages, parts per million, mg/l, mmol/mol or similar; typically representing portions of a total sample weight or volume. Some examples include multivariate measurements of pollutant concentrations, water chemistry, air volatile compounds, foodstuff nutritional compositions, or species relative abundances. They can be entered in an explanatory role in a regression problem, for instance to assess their relationship with a water, air or food quality index. Compositional variables are also common in social sciences like economics. For example, shares of enterprise size classes in a region, investment portfolios, and household or time budgets; which may be put in relation to a productivity or profitability indicator. Such variables carrying relative information are regarded as intrinsically interrelated parts of a so-called *composition*, and their observations are generally referred to as *compositional data* (Aitchison, 1986). Proper statistical processing of compositional data (i.e., accounting for their specific nature) is a key requirement for obtaining interpretable results, but also contributes to the overall validity of the statistical analysis (Pawlowsky-Glahn et al., 2015; Filzmoser et al., 2018). This also holds if the compositional parts act as covariates in regression analysis (Hron et al., 2012).

In practice, a common issue is that the observed data set contains *outliers*, i.e. observations that deviate from the majority of the observations. This can occur for different reasons, including measurement error or some form of contamination. Unfortunately, outliers can greatly influence estimates of the model parameters and may lead to unreliable results. Methods have been developed in the literature to downplay the effects of outliers in order to make statistical analysis more robust (see, e.g., Hampel et al., 1986; Rousseeuw and Leroy, 1987; Maronna et al., 2002; Huber and Ronchetti, 2009). Traditionally, robust estimators for multivariate data have been designed to deal with entire observations being contaminated, assuming that there is a majority of non-contaminated observations in the data set. Such outliers are in the following referred to as *rowwise outliers*, in reference to the fact that observations are commonly arranged by rows in a data matrix, whereas the variables of interest are arranged by columns. However, atypical observations often exhibit outlying values only in a single variable or a small subset of variables (Rousseeuw and Van den Bossche, 2018). When contamination occurs at the cell level of a data matrix, it is actually possible that the majority of rows contain some outlying cells. Thus, treating entire observations as outliers might lead to an unacceptable loss of useful information, particularly in high-dimensional data sets. In the literature on rowwise outliers, equivariance properties are considered essential for estimators, and robustness properties such as the breakdown point are linked to equivariance properties (e.g., Lopuhaä and Rousseeuw, 1991). For robustness against outlying cells, on the other hand, it is necessary to give up properties such as affine equivariance, as affine transformations can spread an outlying cell over all components of the observation (Alqallaf et al., 2009). Recent literature has focused on this latter type of outliers, referred to as *cellwise outliers*, although this literature is still scarce. Some examples include works addressing outlier detection (Rousseeuw and Van den Bossche, 2018), scatter matrix estimation (e.g., Van Aelst et al., 2011; Agostinelli et al., 2015; Leung et al., 2017), linear regression (e.g., Öllerer et al., 2016; Leung et al., 2016; Filzmoser et al., 2020), principal component analysis (e.g., Hubert et al., 2019), and clustering (e.g., Farcomeni, 2014a,b). Figure 1 illustrates the two types of outliers that can be found in a data matrix. In addition to issues with outliers, when

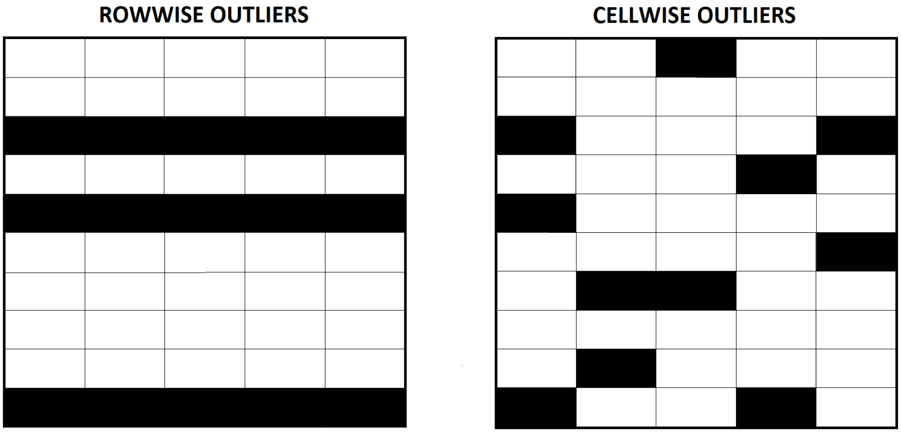

Fig. 1: Illustration of rowwise outliers (*left*) and cellwise outliers (*right*).

working with compositional data we have to take into account that all the relevant information about a compositional part is contained in the ratios between parts (Pawlowsky-Glahn et al., 2015).

In this paper, we introduce a robust estimation procedure for regression analysis with compositional covariates that is designed to handle both cellwise and rowwise outliers. The key idea is to first detect outlying cells and subsequently replace them by sensible values using a (rowwise) robust imputation procedure. Our simulations indicate that when only a few cells of a row are contaminated, treating outliers at the cell level with the proposed procedure (rather than at the row level with rowwise-only robust compositional regression) is advantageous even when the number of explanatory variables is relatively small (see Section 4). This is particularly relevant in the presence of a complex (compositional) data structure, because the pernicious effects of cellwise outliers easily propagate through the ratios between compositional parts. Nevertheless, as the two types of outliers may occur simultaneously in a data set (Leung et al., 2016; Rousseeuw and Van den Bossche, 2018), it is important to note that our method is able to protect against both cellwise and rowwise outliers.

The proposed robust procedure is developed for a linear regression model with a real-valued response and compositional explanatory variables, possibly accompanied by additional real-valued covariates. It is similar in spirit to the 3-step regression estimator of Leung et al. (2016). Both methods start by filtering cellwise outliers and then apply rowwise robust methods. As Leung et al. (2016) only consider real-valued variables, they can use a rowwise robust estimator for incomplete data (see also Danilov et al., 2012). However, the situation is more challenging with compositional data, as they need to be represented in an appropriate coordinate system for proper statistical analysis. This is not feasible with incomplete data (see Appendix B), which is why our procedure makes use of an imputation step. The shooting S-estimator of Öllerer et al. (2016), on the other hand, takes a very different approach. It does not contain a filtering step, but instead combines a coordinate descent algorithm with simple robust regressions to handle outlying cells. Most coordinate representations for compositional data are therefore not suitable for the shooting S-estimator due to the propagation of outliers.

In Section 2, we provide some statistical background about compositional data analysis and introduce the particular logratio coordinate system we use to represent compositional variables in a regression context. We focus on so-called pivot coordinates, which allow to link each compositional part to a logratio coordinate within an orthonormal coordinate system. Specifically, such a logratio coordinate isolates all the relative information about the corresponding compositional part with respect to the other parts in a given composition. Pivot coordinates have been successfully used in regression analysis with compositional covariates (Hron et al., 2012), as well as in regression-based imputation of missing values in compositional data (Hron et al., 2010). Unlike Hron et al. (2012), we perform regression with the MM-estimator (Yohai, 1987) to achieve high robustness with tunable efficiency, but any other rowwise robust regression method could be used in our procedure instead. Section 3 gives a detailed description of the proposed method, which is designed for the regular case of regression analysis with more observations than explanatory variables. Its relative performance in comparison to other regression methods is assessed by simulation in Section 4, whereas Section 5 illustrates its use in a bio-environmental science application. The results indicate that our procedure, which maximizes the use of the information contained in the data set, can cope with moderate levels of cellwise and rowwise contamination, and yields better or comparable estimates than its competitors: the aforementioned 3-step regression estimator and shooting S-estimator, as well as the rowwise robust MM-estimator and the ordinary least squares estimator. Moreover, our procedure allows to perform regression analysis in any isometric logratio coordinate system that provides suitable interpretability of the results, whereas the predicted values do not depend on the particular coordinate representation. Section 6 compares our procedure to its competitors in terms of computation time, and the final Section 7 concludes.

2 Methodological background

A D-part composition is defined as a random vector $\boldsymbol{X} = (X_1, \ldots, X_D)'$ with strictly positive components (compositional parts), carrying relative information. Accordingly, compositional data are multivariate observations where the relevant information is contained in the ratios between parts (Pawlowsky-Glahn et al., 2015). Compositions are commonly represented as proportions or percentages (where the sum of the parts is equal to 1 or 100, respectively). However, the above definition implies that the sample space is actually formed by equivalence classes of proportional vectors and the particular value of the sum of the compositional parts is irrelevant. Instead of ratios, it is advantageous to work with logratios when dealing with compositions, as logratios map the range of a ratio from the positive real space onto the entire real space and symmetrize their values around zero. Moreover, inverse logratios provide the same information up to the sign, i.e., $\ln(X_j/X_k) = -\ln(X_k/X_j)$. This relationship implies that for the purpose of cellwise outlier detection, only $D(D-1)/2$ instead of D^2 logratios have to be considered.

Let $\mathbf{x} = (x_1, \ldots, x_D)'$ be an observation of a random composition $\boldsymbol{X} = (X_1, \ldots, X_D)'$. Clearly, if a form of contamination generates an outlying value in a compositional part x_j, this will affect all pairwise logratios where x_j is contained. On the other hand, data contamination that generates just one aberrant pairwise logratio $\ln(x_j/x_k)$ might have been originated from two outlying compositional parts, namely x_j and x_k. These considerations need to be taken into account when developing a cellwise outlier detection method in the context of compositional data analysis.

Compositional data formally obey the so-called Aitchison geometry of the simplex sample space (Pawlowsky-Glahn et al., 2015). Therefore, it is necessary to map compositions onto the real space in order to apply ordinary statistical methods that rely on the real Euclidean geometry. From a geometrical perspective, a new coordinate system with respect to the Aitchison geometry is constructed. For our purpose, so-called isometric logratio (ilr) coordinates are preferable as they allow to express compositions in an orthonormal coordinate system (Egozcue et al., 2003).

Accordingly, the ilr mapping is such that distances between points in the original Aitchison geometry of the simplex are preserved in the real Euclidean geometry of $\mathbb{R}^{D-1}$. Specifically, we choose so-called pivot coordinates where the role of a single compositional part against the others is highlighted (Fišerová and Hron, 2011; Hron et al., 2017). This way, for a D-part composition $\boldsymbol{X} = (X_1, \ldots, X_D)'$, we obtain a real-valued random vector $\boldsymbol{Z} = (Z_1, \ldots, Z_{D-1})'$ with

$$Z_j = \sqrt{\frac{D-j}{D-j+1}} \ln \frac{X_j}{\sqrt[D-j]{\prod_{k=j+1}^{D} X_k}}, \qquad j = 1, \ldots, D-1. \tag{1}$$

Thus, all relative information about X_1 — with respect to the (geometric) average of the remaining parts — is contained in the first coordinate Z_1. Equivalently, Z_1 can be expressed as

$$Z_1 = \frac{1}{\sqrt{D(D-1)}} \left[\ln(X_1/X_2) + \cdots + \ln(X_1/X_D)\right], \tag{2}$$

i.e., as a (scaled) sum of all the pairwise logratios with X_1 in the numerator. By permuting the compositional parts in $\boldsymbol{X}$, such that a different part is put at the first position each time, we can obtain D different orthonormal coordinate systems, which are orthogonal rotations of each other. Each of them emphasizes the role of the respective compositional part placed at the first position (Fišerová and Hron, 2011). We then generalize the expression in (1) by denoting $\boldsymbol{X}^{(l)} = \left(X_1^{(l)}, \ldots, X_D^{(l)}\right)' = (X_l, X_2, \ldots, X_{l-1}, X_{l+1}, \ldots, X_D)'$ and $\boldsymbol{Z}^{(l)} = \left(Z_1^{(l)}, \ldots, Z_{D-1}^{(l)}\right)'$, with

$$Z_j^{(l)} = \sqrt{\frac{D-j}{D-j+1}} \ln \frac{X_j^{(l)}}{\sqrt[D-j]{\prod_{k=j+1}^{D} X_k^{(l)}}}, \quad j = 1, \ldots, D-1, \quad l = 1, \ldots, D. \tag{3}$$

Thus, all the relative information about an arbitrary compositional part X_l, $l = 1, \ldots, D$, is contained in the corresponding first pivot coordinate $Z_1^{(l)}$. Note that an inverse mapping can be applied to transform back to $\boldsymbol{X}^{(l)} = \left(X_1^{(l)}, \ldots, X_D^{(l)}\right)'$, with

$$\begin{aligned}
X_1^{(l)} &= \exp\left(\sqrt{\frac{D-1}{D}} Z_1^{(l)}\right), \\
X_j^{(l)} &= \exp\left(-\sum_{k=1}^{j-1} \frac{1}{\sqrt{(D-k+1)(D-k)}} Z_k^{(l)} + \sqrt{\frac{D-j}{D-j+1}} Z_j^{(l)}\right), \\
&\qquad\qquad\qquad\qquad\qquad\qquad\qquad\qquad j = 2, \ldots, D-1, \\
X_D^{(l)} &= \exp\left(-\sum_{k=1}^{D-1} \frac{1}{\sqrt{(D-k+1)(D-k)}} Z_k^{(l)}\right).
\end{aligned} \tag{4}$$

This conveniently allows to transfer outputs from statistical processing in real space back to the original simplex sample space of compositional data, using any possible proportional representation within the equivalence class according to any prescribed sum of parts (Filzmoser et al., 2018).

3 Robust compositional regression with cellwise outliers

Here we address three challenges for regression analysis: (i) the inclusion of compositional explanatory variables, possibly complemented by real-valued explanatory variables; (ii) the presence of cellwise outliers; and (iii) the presence of rowwise outliers. Each one creates its own set of particular issues for statistical modelling, and regardless of their occurrence in isolation or in combination, ignoring these issues can lead to unreliable and biased results (e.g., Hron et al., 2012; Filzmoser et al., 2018; Öllerer et al., 2016; Leung et al., 2016; Yohai, 1987). Therefore, the proposed method consists of four stages:

1. Detect outlying cells in the data set (that are not part of entire outlying observations).
2. Replace them by sensible values via rowwise robust imputation (possibly in both response and covariates).
3. Conduct rowwise robust regression using the imputed data, including compositional predictors conveniently expressed in terms of logratio pivot coordinates.
4. Use a multiple imputation scheme so that the standard errors of the regression coefficient estimates account for the additional uncertainty caused by missing values.

These stages are discussed in more detail in the following subsections, while pseudocode for the entire procedure is given in Appendix A. Note that the separate imputation step is necessary to keep the properties of the pivot coordinates defined in (3). If the filtered outliers were not imputed, the resulting missing logratios of compositional parts would make it unmanageable to work with pivot coordinates. As an alternative, we investigated a modification of pivot coordinates so that this propagation of missing values is avoided (see Appendix B). However, these modified pivot coordinates do not lead to coordinate systems that are exact orthogonal rotations of each other, and therefore their practical interpretability is compromised. It should also be noted that we include the response variable in the cellwise outlier filter and multiple imputation steps, which is in line with the literature on multiple imputation (e.g., Little, 1992; Allison, 2002, p.53). Omitting the possibly correlated response variable from the imputation models would in general imply misspecification of the conditional distributions from which the imputed values are drawn, yielding biased estimates of the regression coefficients (see Appendix C). If a prediction of the response variable is needed for a new observation that contains missing values in the explanatory variables, a separate imputation procedure that considers only the explanatory variables could be applied before predicting the response.

For the sake of easing the description of the data set involved in each of the four stages and the corresponding roles of the variables, we use the following mathematical notation:

- $R_1, \dots, R_p, R_{p+1}$ to indistinctly refer to any potential real-valued covariates $V_1, \dots, V_p$ along with the response variable Y, whenever their distinction is not relevant.
- $\boldsymbol{\mathcal{X}} = (\boldsymbol{x}_1, \dots, \boldsymbol{x}_D, \boldsymbol{r}_1, \dots, \boldsymbol{r}_{p+1})$ to represent an $n \times (D + p + 1)$ dimensional data matrix in which the rows contain realizations of the compositional parts $X_1, \dots, X_D$ and the real-valued variables $R_1, \dots, R_{p+1}$, with $\boldsymbol{x}_j = (x_{1j}, \dots, x_{nj})'$ and $\boldsymbol{r}_j = (r_{1j}, \dots, r_{nj})'$ representing column vectors of observations of each of them. The corresponding imputed data set is denoted by $\tilde{\boldsymbol{\mathcal{X}}}$, and its compositional and real-valued elements are denoted by $\tilde{x}_{i1}, \dots, \tilde{x}_{iD}$ and $\tilde{r}_{i1}, \dots, \tilde{r}_{i,p+1}$, respectively, $i = 1, \dots, n$.
- $\boldsymbol{\mathcal{L}} = (\ln(\boldsymbol{x}_1/\boldsymbol{x}_2), \dots, \ln(\boldsymbol{x}_{D-1}/\boldsymbol{x}_D), \boldsymbol{r}_1, \dots, \boldsymbol{r}_{p+1})$ to refer to an $n \times [D(D-1)/2+p+1]$ dimensional data matrix in which the compositional parts are represented by all the corresponding $D(D-1)/2$ pairwise logratios.
- $\mathcal{O}$ for the set of indices (i, j) of the cells in $\boldsymbol{\mathcal{X}}$ that are marked as cellwise outliers. It is important to recall that outlying cells are only marked if they are not part of a rowwise outlier.

3.1 Detection of cellwise outliers

The detection of deviating cells is based on the bivariate filter of Rousseeuw and Van den Bossche (2018). The foremost assumption of this method is that the data matrix is generated from a multivariate normal population, but some cell values are contaminated at random and become outliers. The procedure is briefly sketched in the following (see Rousseeuw and Van den Bossche, 2018, for full details):

1. First, all variables (columns) are robustly standardized, e.g., by subtracting the median and dividing by the median absolute deviation (MAD).
2. Then deviating cells in single variables are marked, i.e., those containing absolute values higher than the cut-off value $\sqrt{\chi^2_{1,\tau}}$, where $\chi^2_{1,\tau}$ is the τ-quantile of the χ^2 distribution with one degree of freedom.
3. For each variable, the correlated variables are determined, i.e., those with absolute robust correlation higher than 0.5. Predictions for every cell are made based on each correlated variable that has a nonmarked cell in the same observation (row). If multiple nonmarked cells are available, the weighted mean of the corresponding predictions can be taken as the predicted value (see Equation (9) of Rousseeuw and Van den Bossche, 2018). A deshrinkage step is subsequently applied to obtain the final prediction. If all other cells of the row are marked as well, the prediction is set to 0 (which is the location estimate of the variable since all variables are standardized). A cell for which the observed value differs too much from its prediction is marked.
4. The cells marked in step 2 or 3 are considered to be cellwise outliers.
5. Finally, rowwise outliers are identified. The i-th row of the data matrix is marked as an outlier if the absolute value of a robustly standardized statistic T_i exceeds the cut-off value $\sqrt{\chi^2_{1,\tau}}$. The statistic T_i is defined as the average (over j) of $F(d^2_{ij})$, where F stands for the cumulative distribution function of the χ^2 distribution with one degree of freedom, and d_{ij} denotes the robustly standardized difference between the value in the cell with indices (i,j) and its prediction (from step 3).

We apply the bivariate filter to the data matrix $\boldsymbol{\mathcal{L}}$, which contains the relevant pairwise logratios of the compositions along with potential real-valued covariates and the response variable, i.e., $\boldsymbol{\mathcal{L}} = (\ln(\boldsymbol{x}_1/\boldsymbol{x}_2), \ldots, \ln(\boldsymbol{x}_{D-1}/\boldsymbol{x}_D), \boldsymbol{r}_1, \ldots, \boldsymbol{r}_{p+1})$. The next task is to transfer the information about the cellwise outliers in $\boldsymbol{\mathcal{L}}$ to $\boldsymbol{\mathcal{X}} = (\boldsymbol{x}_1, \ldots, \boldsymbol{x}_D, \boldsymbol{r}_1, \ldots, \boldsymbol{r}_{p+1})$. While this is identical for the real-valued variables, we propose to mark a compositional part x_{ij} in $\boldsymbol{\mathcal{X}}$ as a cellwise outlier (and subsequently set its value to missing to be imputed) if at least half of the logratios containing x_{ij} are identified as outliers by the bivariate filter. After extensive simulation experiments, we found this condition strict enough to detect outlying compositional parts but not overly strict. As a matter of fact, many outlying cells would not be detected if we required that all logratios including a particular part had to be marked as outliers. Rousseeuw and Van den Bossche (2018) recommend to use $\tau = 0.99$ in the cut-off value $\sqrt{\chi^2_{1,\tau}}$ of the outlier filter, which gave favorable results in our simulations. Nevertheless, we recommend to consider lower values of τ as well to investigate sensitivity relative to this parameter (as illustrated in the case study in Section 5).

Note that the purpose of the initial filter is to avoid that the subsequent regression modelling is influenced by cellwise outliers. However, while cellwise outlier filters perform well in detecting individual outlying cells, they are not as effective in detecting rowwise outliers (Leung et al., 2016; Rousseeuw and Van den Bossche, 2018). Hence it is still crucial to protect against rowwise outliers in the subsequent stages of the procedure. Moreover, observations that have a large number of outlying cells are likely to be rowwise outliers. In our view, it is thus better not to impute those data cells and instead have the entire observation downweighted by a robust regression estimator in the following stages. Hence, at this point we treat an observation as a rowwise outlier if step 5 of the bivariate filter identifies the corresponding row in $\mathcal{L}$ as a rowwise outlier, or if at least 75% cells of the corresponding row in $\boldsymbol{\mathcal{X}}$ are marked as cellwise outliers. The final index set $\mathcal{O}$ contains the indices (i, j) of all cellwise outliers that are not part of rowwise outliers. Cells of $\boldsymbol{\mathcal{X}}$ indicated by $\mathcal{O}$ are treated as missing values to be imputed in the next stage.

3.2 Imputation of cellwise outliers

Since compositional data are projected onto $\mathbb{R}^{D-1}$ through logratios involving several parts, missing parts as derived from the cellwise outlier filter can easily result in an unmanageable amount of missing logratios. We therefore impute the affected cells beforehand, so that subsequent compositional regression based on logratios can be conducted as usual on the imputed data matrix. For this purpose, we modify the iterative model-based imputation procedure of Hron et al. (2010) for compositional data to allow for a mixture of compositional and real-valued variables. This method uses a representation of the compositional data in pivot coordinates, and imputes the missing cells by estimates of expected values conditional on the observed part of the data. Such conditional expected values are modeled by linear regression models (with the assumption that the error terms have expected value equal to zero), which are fitted using the rowwise robust MM-estimator (Yohai, 1987). As MM-regression allows to reduce the influence of rowwise outliers on the estimation of the imputation model, the imputed values may reflect the structure of the majority of the available data.

The imputation of outlying cells starts by separately sorting compositional parts and real-valued variables in decreasing order according to the amount of missing values. To simplify notation, we assume that this sorting does not change the original position of any compositional part or real-valued variable.

Following Hron et al. (2010), the imputation algorithm is initialized with the simultaneous k-nearest-neighbor ($k$nn) method, which is based on the Aitchison distance (Pawlowsky-Glahn et al., 2015) between neighbors for the compositional parts and on the Euclidean distance between neighbors for the real-valued variables.

Each iteration of the imputation algorithm consists of at most $D + p + 1$ steps. The first steps involve the imputation of the compositional parts (up to D), whereas the remaining steps involve the imputation of the real-valued variables (up to $p + 1$). The procedure is summarized as follows:

1. For each compositional part $\boldsymbol{x}_l$ that contains outlying cells, $l = 1, \ldots, D$, pivot coordinates are obtained to sequentially fit regression models of $Z_1^{(l)}$ on the remaining $D - 2$ coordinates plus the $p + 1$ non-compositional variables as covariates:
$$Z_1^{(l)} = a + b_2^{(l)} Z_2^{(l)} + \ldots + b_{D-1}^{(l)} Z_{D-1}^{(l)} + c_1 R_1 + \ldots + c_{p+1} R_{p+1} + \varepsilon^{(l)}, \tag{5}$$
where $\varepsilon^{(l)}$ is a random error term. Observations with no outlying cell in $\boldsymbol{x}_l$ are used for model fitting. The estimated regression coefficients $\hat{a}$, $\hat{b}_2^{(l)}, \ldots, \hat{b}_{D-1}^{(l)}, \hat{c}_1, \ldots, \hat{c}_{p+1}$ are obtained using MM estimation such that they are robust against rowwise outliers. Furthermore, MM-regression also protects against poorly initialized missing value imputation (Hron et al., 2010). The coefficient estimates are then used to compute predicted values $\hat{z}_{i1}^{(l)}, (i, l) \in \mathcal{O}$.
For $(i, l) \in \mathcal{O}$, imputed compositional parts $\hat{x}_{i1}, \ldots, \hat{x}_{iD}$ are obtained from the pivot coordinates $\hat{z}_{i1}^{(l)}, z_{i2}^{(l)}, \ldots, z_{i,D-1}^{(l)}$ via the inverse mapping in (4). Note that the ratios between the non-outlying parts are not affected by this procedure.

2. Next, each real-valued variable that contains outlying cells is imputed in an analogous way by sequentially serving as response in MM-regression on the remaining variables as predictors, including the compositional parts through pivot coordinates. Note that it does not matter which particular pivot coordinate system is used here. They all yield the same predictions due to the fact that they are orthogonal rotations of each other.

This is repeated iteratively until the sum of the squared relative changes in the imputed values are smaller than a threshold η. Following Hron et al. (2010), η was set at 0.5, and only a few iterations were typically needed to reach convergence in our simulations. This results in an imputed data set $\tilde{\boldsymbol{\mathcal{X}}}$ which serves as input for the subsequent stage.

The performance of the imputations in the steps 1 and 3 above can often be improved by applying some form of variable selection to fit the corresponding regression models. To keep the computational burden low, we use a simple initial variable screening technique: before starting the iterative imputation procedure, we identify the most correlated variables for each variable to be imputed. We thereby compute robust correlations via bivariate winsorization (Khan et al., 2007) based on pairwise complete observations. However, initial simulations suggest that variable screening may not be necessary if the number of variables and the amount of filtered cells are both relatively small (e.g., $D + p + 1 \leq 10$ and less than 10% filtered cells). Moreover, when the number of variables is small, a smaller correlation threshold should be used to ensure that enough variables survive the screening process. Our procedure therefore implements the following default behavior as a compromise: if $D + p + 1 \leq 10$, only variables with absolute correlations higher than 0.2 are used, otherwise the threshold is set to 0.5.

3.3 Robust compositional regression

After imputing cellwise outliers, and possibly other missing values in the data set, the actual regression modelling is conducted. Hron et al. (2012) proposed a suitable approach for regression with compositional explanatory variables that yields a meaningful interpretation of the regression coefficients through the use of logratio pivot coordinates. Here we extend this to include the possibility of having p non-compositional covariates $V_1, \ldots, V_p$ along with the D-part composition $\boldsymbol{X} = (X_1, \ldots, X_D)'$ as predictors of a real-valued response variable Y. By expressing the composition in $\mathbb{R}^{D-1}$ via pivot coordinates defined in (3), we obtain D different linear regression models

$$Y = \alpha + \beta_1^{(l)} Z_1^{(l)} + \ldots + \beta_{D-1}^{(l)} Z_{D-1}^{(l)} + \gamma_1 V_1 + \ldots + \gamma_p V_p + \varepsilon, \quad l = 1, \ldots, D, \tag{6}$$

with regression parameters $(\alpha, \beta_1^{(l)}, \ldots, \beta_{D-1}^{(l)}, \gamma_1, \ldots, \gamma_p)'$, $l = 1, \ldots, D$, and a random error term ε. Parameter estimation is conducted by ordinary least squares in Hron et al. (2012). But since we still need to protect against rowwise outliers after dealing with cellwise outliers, we instead apply the robust and highly efficient MM-estimator (Yohai, 1987). Note that this estimator is designed to handle rowwise outliers only, and it could easily fail if applied directly to data containing cellwise outliers by skipping the previous cellwise outlier detection and imputation stages. The same problem would occur with other rowwise robust estimators for regression models with compositional data (e.g., Hron and Filzmoser, 2010; Hrůzová et al., 2016).

As $Z_1^{(l)}, \ldots, Z_{D-1}^{(l)}$ for different choices of l result from orthogonal rotations of the corresponding pivot coordinate systems, the associated regression fits yield identical estimates of the intercept and the regression coefficients of the non-compositional covariates, which are denoted by $\hat{\alpha}$ and $\hat{\gamma}_1, \ldots, \hat{\gamma}_p$, respectively. Moreover, the (normalized) aggregation of all pairwise logratios involving X_l into the coordinate $Z_1^{(l)}$ results in a logratio that stands for the *dominance* of the

l-th part with respect to the average of the other components (in terms of the geometric mean, see (3)). Accordingly, the value of the coefficient $\beta_1^{(l)}$ relates to the influence of the dominance of the part X_l (with respect to the mean behavior of the other parts in the composition) on the response variable. Because of the mutual orthogonality of the pivot coordinate systems, we can sequentially extract the estimate $\hat{\beta}_1^{(l)}$ from each of the D models fitted above ($l = 1, \ldots, D$). Hence, the final vector of estimated regression coefficients is $(\hat{\alpha}, \hat{\beta}_1^{(1)}, \ldots, \hat{\beta}_1^{(D)}, \hat{\gamma}_1, \ldots, \hat{\gamma}_p)'$.

Following Müller et al. (2018), the interpretation of the coefficients of the compositional parts can be enhanced by ignoring the normalization constant of the respective pivot coordinate in (3) and using binary logarithms rather than natural logarithms. This way, doubling the dominance of X_l implies a unitary increase of the binary logarithm. Accordingly, under the usual assumption that the error terms of the model have expected value equal to zero, the value of the coefficient $\beta_1^{(l)}$ corresponds to the change in the mean response when the dominance of X_l is doubled, while keeping all other regressors fixed. Nevertheless, we apply the normalization constant and use natural logarithms (as commonly done) for the purpose of this paper.

3.4 Multiple imputation estimates

As described above, the input data to fit the final regression model is an imputed data set $\tilde{\boldsymbol{\mathcal{X}}}$. It is well-known that measures of variability like standard errors can be underestimated when the usual formulas are applied to imputed data (Little and Rubin, 2002). Consequently, statistical significance tests in relation to the regression coefficients tend to be anticonservative. The reason is that the uncertainty derived from imputing the filtered cells is not taken into account. A well-established solution to this problem is using multiple imputation (MI) (Rubin and Schenker, 1986). The basic idea is that instead of a single imputed data set, M different imputed data sets are actually analysed. It has been shown that by aggregating estimates from all these data sets, better estimates of the standard errors are obtained, as they reflect the additional uncertainty from the imputation process (Little and Rubin, 2002; Van Buuren, 2012; Cevallos Valdiviezo and Van Aelst, 2015). We adopt this approach and, following Bodner (2009) and White et al. (2011), we consider the number of imputed data sets M to be the rounded percentage of rows in the data matrix affected by cellwise outliers.

Each of the M data sets is obtained from $\tilde{\boldsymbol{\mathcal{X}}}$ by adding random noise to the estimated values resulting from the imputation procedure (Section 3.2). That is, rather than imputing the filtered cells with the conditional expected value, we impute them by a random draw from the estimated conditional distribution. For compositional data, the noise is not added directly to the compositional part $\tilde{x}_{il}, (i,l) \in \mathcal{O}$, as this would be incoherent with the geometry of the simplex, but to the first pivot coordinate $\tilde{z}_{i1}^{(l)}$, obtained from the composition $\left(\tilde{x}_{i1}^{(l)}, \ldots, \tilde{x}_{iD}^{(l)}\right)' = (\tilde{x}_{il}, \tilde{x}_{i1}, \ldots, \tilde{x}_{i,l-1}, \tilde{x}_{i,l+1}, \ldots, \tilde{x}_{iD})'$ via (3). The corresponding values of the compositional parts are then obtained by the inverse mapping in (4). More specifically, consider the j-th step of the last iteration of the imputation procedure (Section 3.2), with $j = 1, \ldots, D+p+1$. Missing values in the j-th variable are imputed by robust regression using all the other variables as predictors. Following Templ et al. (2011b), random noise is added to the imputed value by drawing M random values from $N(0, \hat{\sigma}_j^2(1 + o_j/n))$, where $\hat{\sigma}_j$ is a robust residual scale estimate from the corresponding regression fit and o_j denotes the number of values to be imputed in the j-th variable.

Afterwards, robust MM-regression estimation (Section 3.3) is performed for each of the M imputed data sets. Following Rubin (1987) and Barnard and Rubin (1999), we use $\hat{\theta}^{\{m\}}$ to denote generically a parameter point estimate (i.e., any of $\hat{\alpha}, \hat{\beta}_1^{(1)}, \ldots, \hat{\beta}_1^{(D)}, \hat{\gamma}_1, \ldots, \hat{\gamma}_p$) and $\hat{U}^{\{m\}}$ refers to the corresponding estimated variance from the m-th imputed data set, $m = 1, \ldots, M$. A final point estimate and variance for each regression coefficient is then obtained as

$$\hat{\theta} = \frac{1}{M} \sum_{m=1}^{M} \hat{\theta}^{\{m\}} \qquad \text{and} \qquad \hat{V} = \hat{W} + \frac{M+1}{M} \hat{B},$$

respectively, where $\hat{W} = \frac{1}{M} \sum_{m=1}^{M} \hat{U}^{\{m\}}$ is the average within-imputation variance and $\hat{B} = \frac{1}{M-1} \sum_{m=1}^{M} \left(\hat{\theta}^{\{m\}} - \hat{\theta} \right)^2$ is the between-imputation variance.

4 Simulation study

In order to assess the performance of our procedure in comparison to other (robust) methods for compositional regression, we perform a simulation study.

4.1 Simulation design

The parameters for the simulation design are partly inspired by the data set about livestock methane emission from ruminal volatile fatty acids (VFA) introduced in Section 5. As the main novelty of our procedure is the inclusion of compositional covariates in the context of robust regression with cellwise and rowwise outliers, we assume for simplicity that there are only compositional covariates involved. We set $n \in \{50, 100, 200\}$ as the number of observations and $D \in \{5, 10, 20\}$ as the number of compositional parts. The simulated compositions are generated through pivot coordinates. In order to obtain a realistic covariance structure in the pivot coordinate system, we chose an initial covariance matrix $\boldsymbol{\Sigma}_0 = \left(0.5^{|i-j|}/10\right)_{1 \leq i,j \leq D-1}$, with entries being similar in magnitude to the ones observed in the VFA case study. To investigate the effects of adding more variability to the data matrix, we consider the covariance matrix in pivot coordinates $\boldsymbol{\Sigma}$ as a multiple of the initial covariance matrix, i.e., $\boldsymbol{\Sigma} = k\boldsymbol{\Sigma}_0$ with $k \in \{1, 2, 3\}$.

We examine a scenario with both rowwise and cellwise outliers. Specifically, we consider the case where outlying rows (entire observations) and outlying cells (in the compositional parts and the response variable) both occur with probability $\zeta \in \{0, 0.02, 0.05, 0.1, 0.2\}$. We first generate entire outlying observations (rows) and, subsequently, outlying cells only in non-outlying rows. We perform 1000 simulation runs for each configuration. In each simulation run, the data are generated as follows:

1. Pivot coordinates are sampled as $\mathbf{z}_i = (z_{i1}, \ldots, z_{i,D-1})' \sim \mathcal{N}_{D-1}(\mathbf{0}, \boldsymbol{\Sigma})$, $i = 1, \ldots, n$.
2. The values of the response variable are obtained in the pivot coordinate system as

$$y_i = \beta_0 + \beta_1 z_{i1} + \ldots + \beta_{D-1} z_{i,D-1} + \varepsilon_i, \qquad \varepsilon_i \sim \mathcal{N}(0, 0.25^2), \qquad i = 1, \ldots, n,$$

   with regression parameters $\beta_0 = 0$ and $(\beta_1, \ldots, \beta_{D-1})' = (1, 0, 1, 0, \ldots)'$. The variance of the error terms ε_i is chosen to roughly mimic the signal-to-noise ratio observed in the VFA data.
3. The pivot coordinates $\mathbf{z}_i = (z_{i1}, \ldots, z_{i,D-1})'$ are transformed according to (4) to obtain the corresponding compositions $\mathbf{x}_i = (x_{i1}, \ldots, x_{iD})'$, $i = 1, \ldots, n$.
4. Observations are randomly selected with probability ζ to be turned into rowwise outliers. We first generate outliers in the pivot coordinates along the smallest principal component. Let $\mathcal{U} \subseteq \{1, \ldots, n\}$ denote the set of indices of the rowwise outliers, and let $\mathbf{q}_i = (q_{i1}, \ldots, q_{i,D-1})'$ denote the principal component scores corresponding to $\mathbf{z}_i$. For $i \in \mathcal{U}$, we change the value

of the last component $q^*_{i,D-1} = q_{i,D-1} + 5\sqrt{k}$. Note that the factor $\sqrt{k}$ ensures that the outlier shift is of the same magnitude for the different scalings of the covariance matrix $\boldsymbol{\Sigma} = k\boldsymbol{\Sigma}_0$. After transforming the scores $\mathbf{q}^*_i = (q_{i1}, \ldots, q_{i,D-2}, q^*_{i,D-1})'$ back to pivot coordinates to obtain outlying $\mathbf{z}^*_i = (z^*_{i1}, \ldots, z^*_{i,D-1})'$, we change the respective values of the response variable to

$$y^*_i = \beta^*_0 + \beta^*_1 z^*_{i1} + \ldots + \beta^*_{D-1} z^*_{i,D-1} + \varepsilon_i, \qquad i \in \mathcal{U},$$

with regression parameters $\beta^*_0 = 0$ and $\beta^*_j = -1$, $j = 1, \ldots, D-1$. Using regression coefficients that are very different to those from clean observations ensures that the rowwise outliers are bad leverage points. Finally, the outlying pivot coordinates $\mathbf{z}^*_i = (z^*_{i1}, \ldots, z^*_{i,D-1})'$ are transformed according to (4) to obtain the corresponding outlying compositions $\mathbf{x}^*_i = (x^*_{i1}, \ldots, x^*_{iD})'$, $i \in \mathcal{U}$.

5. Cells corresponding to non-outlying observations $(x_{i1}, \ldots, x_{iD}, y_i)'$, $i \notin \mathcal{U}$, are randomly selected with probability ζ to be turned into cellwise outliers. Let $\mathcal{O}$ denote the set of indices (i, j) of the outlying cells. For any pair $(i, j) \in \mathcal{O}$, we change the cell value to $x^{**}_{ij} = 10 \cdot x_{ij}$ if $j \in \{1, \ldots, D\}$ or to $y^{**}_i = 10 \cdot y_i$ if $j = D + 1$. The multiplicative factor was chosen to minimize the chance that outlying cells overlap with noise that occurs naturally in the composition or the real-valued response.

The resulting observations with rowwise and cellwise outliers are denoted by $\mathbf{x}^\star_i = (x^\star_{i1}, \ldots, x^\star_{iD})'$ and $y^\star_i$, where

$$x^\star_{ij} = \begin{cases} x^*_{ij}, & \text{if } i \in \mathcal{U}, \\ x^{**}_{ij}, & \text{if } (i,j) \in \mathcal{O}, \\ x_{ij}, & \text{otherwise,} \end{cases} \qquad i = 1, \ldots, n, \;\; j = 1, \ldots, D,$$

and

$$y^\star_i = \begin{cases} y^*_i, & \text{if } i \in \mathcal{U}, \\ y^{**}_i, & \text{if } (i, D+1) \in \mathcal{O}, \\ y_i, & \text{otherwise,} \end{cases} \qquad i = 1, \ldots, n.$$

4.2 Methods, performance measures, and software

Below we give a brief description of the methods that participate in the evaluation, together with the abbreviations we use to refer to them:

LS: ordinary compositional least squares regression (with no treatment for outliers).

MM: robust compositional MM-regression (with no treatment for cellwise outliers).

ShS: shooting S-estimator (Öllerer et al., 2016) obtained from the $D(D-1)/2$ unique pairwise logratios. The shooting S-estimator is designed to cope with cellwise contamination by weighing the components of an observation differently. Note that the results can only be compared in terms of prediction and not in terms of parameter estimation. We used both Tukey's biweight loss function and the skipped Huber loss function: the former yields continuous weights in $[0, 1]$ while the latter leads to binary weights in $\{0, 1\}$ (see Öllerer et al., 2016). We only report the results for Tukey's biweight loss function, as it generally gave better and more stable results than the skipped Huber loss function.

3S: 3-step regression (Leung et al., 2016) fitted to additive logratio (alr) coordinates, i.e. the composition $(X_1, \ldots, X_D)'$ is represented by the real-valued vector of log-ratios $\big(\ln(X_1/X_j), \ldots, \ln(X_{j-1}/X_j), \ln(X_{j+1}/X_j), \ldots, \ln(X_D/X_j)\big)'$, using a part X_j as reference in the denominator (Aitchison, 1986). Note that the use of $D(D-1)/2$ pairwise logratios as covariates is not possible here since the algorithm requires full-ranked data. 3-step regression first uses a consistent univariate filter to eliminate outlying cells; second, it applies a robust estimator of multivariate location and scatter to the filtered data to downplay outlying rows; and third, it computes robust regression coefficients from the previous step. In each simulation run, the reference part X_j is selected randomly. As with the shooting S-estimator, the results are compared only in terms of prediction. It is important to note that the predicted values depend on the choice of X_j in the denominator of the logratios. For example, an outlying value in a cell x_{i1} results in a rowwise outlier in the observation $(\ln(x_{i2}/x_{i1}), \dots, \ln(x_{iD}/x_{i1}))'$, but only in a cellwise outlier in $(\ln(x_{i1}/x_{iD}), \dots, \ln(x_{i,D-1}/x_{iD}))'$. These cases will be handled differently by 3-step regression, yielding different predictions of the response variable. Although this leads to somewhat limited practical applicability, it is still informative to include this approach here in order to compare its general performance.

BF-MI: this is our proposed method which applies the bivariate filter (BF) followed by multiple imputation (MI). Based on preliminary simulations, we set $\tau = 0.99$ (to determine the cutoff value for marking outliers in the bivariate filter). In the imputations, we use the default behavior for variable screening (see Section 3.2). For the MM-estimator, we use Tukey's biweight loss function, with the initial estimator tuned for maximum breakdown point and the final estimator tuned for 95% efficiency.

IF-MI: this represents a hypothetical situation where an ideal filter (IF) is able to perfectly identify all outlying cells (and only those). The remaining steps of our method are afterwards applied using multiple imputation (MI). We use the same settings for variable screening and MM estimation as used for BF-MI. This case is included for benchmarking purposes only, as it is generally unattainable in practice.

Note that all methods except the shooting S-estimator and 3-step regression consider pivot coordinates to represent the compositional covariates. By construction, the shooting S-estimator and the 3-step regression method require the use of pairwise logratios and alr coordinates, respectively.

The performance of the methods is assessed in terms of the mean squared error (MSE) of the coefficient estimates, computed as

$$MSE = \frac{1}{D} \sum_{j=0}^{D-1} (\hat{\beta}_j - \beta_j)^2.$$

Note that in order to reduce the computational burden, a single set of pivot coordinates is used without loss of generality to calculate the MSE of the regression coefficients. Further evaluation is made in terms of prediction error. For this purpose, n additional clean test observations $\mathbf{x}_i^{test}$ and y_i^{test}, $i = 1, \dots, n$, are generated in each simulation run according to steps 1–3 of our data generating process. Note that the number of observations in the test data is the same as in the training data to which the methods are applied. On the test data, the mean squared error of prediction (MSEP) is calculated as

$$MSEP = \frac{1}{n} \sum_{i=1}^{n} (\hat{y}_i^{test} - y_i^{test})^2,$$

where $\hat{y}_i^{test}$ denote the predicted values of y_i^{test}.

All computations were performed using the R environment for statistical computing (R Core Team, 2020), including the packages `cellWise` (Raymaekers et al., 2019), `robCompositions` (Templ et al., 2011a), `robreg3S` (Leung et al., 2015) and the function `shootingS()` obtained from `https://github.com/aalfons/shootingS`. The code for our method is available at `https://github.com/aalfons/lmcrCoda`.

4.3 Simulation results

For different numbers of compositional parts D, Figures 5–7 in Appendix D contain plots of the average MSE against the contamination level ζ for various sample sizes n and scaling factors k of the covariance matrix in pivot coordinates. Similarly, the average MSEP is displayed in Figures 8–10 in Appendix D.

Regarding coefficient estimates, all methods are accurate when there is no contamination ($\zeta = 0$). As contamination increases, OLS is quickly influenced by the outliers, yielding the highest MSE of all methods. The MSE of MM also increases continuously for increasing contamination level, which is expected since MM is only robust to rowwise outliers but not to cellwise outliers. Our proposed method BF-MI is however very accurate for up to 5% contamination and close to the hypothetical IF-MI case using an ideal outlier filter. While the MSE of BF-MI increases for larger contamination levels, it is generally still lower than that of MM, although the difference between the two becomes small as variability in the data increases (increasing k). The MSE of IF-MI remains fairly low for 10% contamination, which indicates that the outlier filtering step is crucial for the performance of our proposed method, but under 20% contamination the MSE of IF-MI increases as well. All in all, the assessment based on MSE suggests that BF-MI offers improved performance over existing techniques for regression analysis with compositional covariates.

As to prediction performance, the results are comparable to the above. OLS in general has the highest MSEP, and BF-MI outperforms MM. In many settings, the MSEP of ShS is comparable to that of BF-MI or somewhat higher, but ShS is unstable if the ratio of n/D is small. Furthermore, ShS cannot be applied for $D = 20$ and $n = 50$ or $n = 100$, since the number of pairwise logratios is larger than the number of observations in those cases. 3S is also similar to BF-MI in terms of MSEP while the contamination level is 5% or lower, but each method is performing slightly better than the other in some settings with higher amounts of contamination. While 3S predicts better for lower values of D when the data are more scattered (higher values of k), BF-MI has lower MSEP for $D = 20$.

Note that we also considered counterparts to IF-MI and BF-MI that use single imputation instead of multiple imputation. The results were very similar. This is actually expected, as the main purpose of multiple imputation is to improve standard errors (Little and Rubin, 2002; Van Buuren, 2012; Cevallos Valdiviezo and Van Aelst, 2015), but there should not be large differences in the point estimates of the coefficients (compared to single imputation). Consequently, the bias component of the MSEP should be similar, and the MSEP can only be improved by reducing the variance in the predictions. In multiple imputation, such a reduction in variance would in turn require to decrease the correlation between predictions based on different imputed data sets. However, when the number of imputed cells is rather small, the predictions based on different imputed data sets are still highly correlated. An improvement in prediction performance via multiple imputation can only be expected for larger fractions of imputed cells (cf. results and recommendations of Cevallos Valdiviezo and Van Aelst, 2015), where the correlation between imputed data sets is sufficiently reduced.

5 Illustrative case study

We apply the proposed compositional MM-regression with a bivariate cellwise outlier filter and multiple imputation (BF-MI algorithm) to investigate the association between livestock methane emissions from individual animals and their ruminal volatile fatty acid (VFA) composition, while accounting for the potential effects of other animal and diet-related covariates. The concentrations of VFA were determined by high-performance liquid chromatography from rumen fluid samples taken using a stomach tube. The quality of the chromatography determines the precision of the measurements, and outlying measurements may be related to unstable baselines, noisy detectors, poor resolution of the components, or errors on the part of the operator in preparing the solution or performing the measurement. The data set consists of $n = 239$ observations originating from the study carried out in Palarea-Albaladejo et al. (2017). It includes the following variables:

- *CH_4*: animal methane yield measured in g/kg DMI using indirect respiration chambers.
- *VFA*: 6-part composition measured in mmol/mol of acetate, propionate, butyrate, isobutyrate, isovalerate and valerate.
- *ME*: diet metabolizable energy measured in MJ/kg DM as estimated from feed composition.
- *DMI*: animal dry matter intake in kg/day.
- *Weight*: animal bodyweight in kg.
- *Diet*: type of diet fed to the animal, either: (a) concentrate diet, based on barley and grains with low forage (< 100 g/kg DM); or (b) mixed diet, including forage (400-600 g/kg DM) along with barley and grains.

All four positive-valued variables in the data set (CH_4, ME, DMI and Weight) are log-transformed and thus mapped into real space to better accommodate model assumptions. Moreover, the data set is split by diet type before the bivariate outlier filter (Section 3.1) is applied separately to each resulting subset of data. Overall, 1.26% of rows are marked as rowwise outliers, while 1.96% of cells in the remaining observations are marked as cellwise outliers. Figure 2 highlights these in each numerical variable, as well as the marked rows, in red color.

Note that both the imputation step (Section 3.2) and the regression step (Section 3.3) of our procedure work with categorical variables in the usual way by including dummy variables. Here we add to the list of covariates a dummy variable $\text{Diet}_{\text{Mixed}}$, which takes the value 1 for mixed diet and 0 for a concentrate diet. The regression model is thus specified as

$$\begin{aligned}\ln(\text{CH}_4) = \alpha + \beta_1^{(l)} Z_1^{(l)} + \ldots + \beta_5^{(l)} Z_5^{(l)} + \\ + \gamma_1 \ln(\text{ME}) + \gamma_2 \ln(\text{DMI}) + \gamma_3 \ln(\text{Weight}) + \delta \text{Diet}_{\text{Mixed}} + \varepsilon,\end{aligned} \tag{7}$$

where $l = 1, \ldots, 6$ indicates the successive pivot coordinate systems and corresponding regression coefficients used to isolate the relative role (dominance) of each of the six parts forming the VFA composition through the first pivot coordinate $Z_1^{(l)}$ in each system (Section 2).

We fit the regression model defined in (7) using ordinary compositional LS estimation, compositional MM estimation and the proposed BF-MI method. For BF-MI, we use $\tau = 0.99$ and skip the variable screening in the imputation step, as the number of variables is rather small and fewer than 2% of cells are filtered. Note that in this application we are interested in an interpretation of the results in terms of pivot coordinates, therefore it is not meaningful to apply other methods such as the shooting S-estimator (Öllerer et al., 2016) or 3-step regression (Leung et al., 2016).

Table 1 displays the results using the three estimation procedures considered. Focusing on the VFA composition, LS estimation does not result in a statistically significant association between the dominance of ruminal acetate and methane yield ($p = 0.127$). The MM-estimator

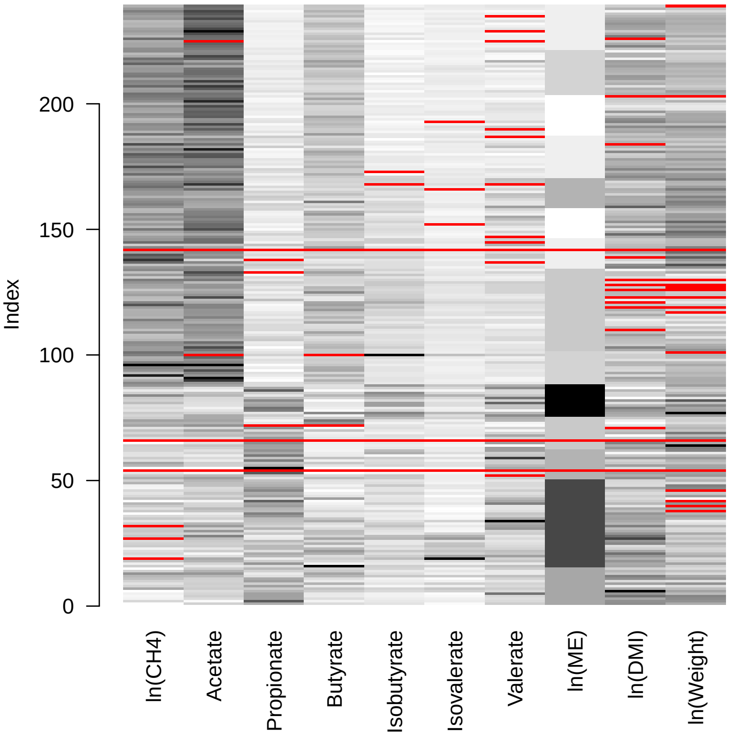

Fig. 2: Cellwise and rowwise outliers detected by the bivariate filter in the VFA data set. Outlying cells/rows are colored in red. The grey color scheme reflects the values of compositional parts and real-valued variables (the higher the value, the darker the color).

Table 1: Regression coefficient estimates, standard errors and p-values for the VFA data set: ordinary compositional LS estimation (LS), compositional MM estimation without a cellwise outlier filter (MM), and proposed compositional MM estimation with a bivariate cellwise outlier filter and multiple imputation (BF-MI).

	LS			MM			BF-MI		
Variable	Coeff.	Std. Error	p-value	Coeff.	Std. Error	p-value	Coeff.	Std. Error	p-value
Intercept	-2.215	1.754	0.208	-3.142	1.965	0.111	-2.822	1.998	0.159
$z_1^{(\text{Acetate})}$	0.125	0.082	0.127	0.203	0.104	0.053	0.301	0.084	< 0.001
$z_1^{(\text{Propionate})}$	-0.247	0.048	< 0.001	-0.304	0.067	< 0.001	-0.385	0.054	< 0.001
$z_1^{(\text{Butyrate})}$	0.093	0.051	0.072	0.070	0.054	0.193	0.025	0.050	0.617
$z_1^{(\text{Isobutyrate})}$	-0.015	0.047	0.744	-0.023	0.052	0.664	-0.014	0.055	0.794
$z_1^{(\text{Isovalerate})}$	0.006	0.032	0.848	0.005	0.034	0.890	0.015	0.034	0.662
$z_1^{(\text{Valerate})}$	0.038	0.039	0.322	0.049	0.037	0.195	0.059	0.064	0.350
ln(ME)	0.725	0.484	0.136	0.999	0.512	0.052	0.755	0.481	0.118
ln(DMI)	-0.413	0.064	< 0.001	-0.408	0.064	< 0.001	-0.397	0.072	< 0.001
ln(Weight)	0.627	0.147	< 0.001	0.651	0.165	< 0.001	0.689	0.186	< 0.001
$\text{Diet}_{\text{Mixed}}$	0.328	0.040	< 0.001	0.308	0.048	< 0.001	0.245	0.048	< 0.001

(without the cellwise outlier filter) provides only a weakly significant positive association between animal methane emission and the relative production of ruminal acetate ($p = 0.053$). Moreover, a statistically significant negative association was concluded in both cases between methane yield and the dominance of propionate ($p < 0.001$). The results from using our proposed BF-MI method are comparable in terms of overall directions of the associations, but the statistical significance of the acetate related term was notably higher ($p < 0.001$), which further stresses the role of the contrast between acetate and propionate as a driver of the association between the ruminal VFA composition and methane emission, which is in agreement with biological knowledge (Wolin, 1960; Palarea-Albaladejo et al., 2017).

As our procedure depends on the parameter τ of the bivariate outlier filter (lower values of τ leading to more cells being marked as cellwise outliers), and on whether variable screening is performed in the imputation step, we perform a sensitive analysis on those parameters. Table 2 in Appendix E shows the results obtained for various sensible choices of τ with and without variable screening. Even though there are some differences in the values of the coefficient estimates, the results are qualitatively similar. The p-values lead to the same conclusions in terms of statistical significance, making the findings robust across all choices.

6 Computation time

We evaluate the computation time of the proposed procedure on simulated data sets. As in Section 4, we vary the number of observations $n \in \{50, 100, 200\}$ and the number of compositional parts $D \in \{5, 10, 20\}$. We follow the same procedure as described in Section 4.1 to generate the data, but only consider $k = 1$ for the multiplicative factor of the covariance matrix, contamination level $\zeta = 0.02$, and 100 simulated data sets for each parameter configuration. For the sake of comparison, we include the same methods as described in Section 4.2. We thereby use the same parameter choices, and the same software packages and functions for their computation. All computation times are measured with the R package `microbenchmark` (Mersmann, 2019) on a laptop with a 2.3 GHz Intel Core i5 processor and 16GB main memory.

The results are shown in Figure 3: the average computation time in seconds is displayed in the top row, and the relative speed gain of each method with respect to our method is displayed in the bottom row. Note that when $D = 20$, the shooting S-estimator (ShS) cannot be applied for $n = 50$ and $n = 100$, as the number of pairwise logratios is larger than the number of observations. Below we give a summary of the relative performance of the other methods with respect to our proposal.

LS: this is about 40–45 times faster than our method with a 5-part composition (depending on the number of observations), and the relative speed difference increases with increasing number of compositional parts.

MM: this is about 25–30 times faster than our method with a 5-part composition (depending on the number of observations), and the relative speed difference increases with increasing number of compositional parts.

ShS: this is about 3–4 times faster than our method, with the relative speed difference being fairly stable in the number of observations and the number of compositional parts.

3S: this is about 5–10 times faster than our method with a 5-part composition (depending on the number of observations). The relative speed difference increases at first with the number of compositional parts, but decreases again when it becomes large enough for our method to perform variable screening in the imputation stage.

There is clearly a price to pay in terms of computation time for the robustness and interpretability of our procedure. However, we find the computation time to be reasonable for many

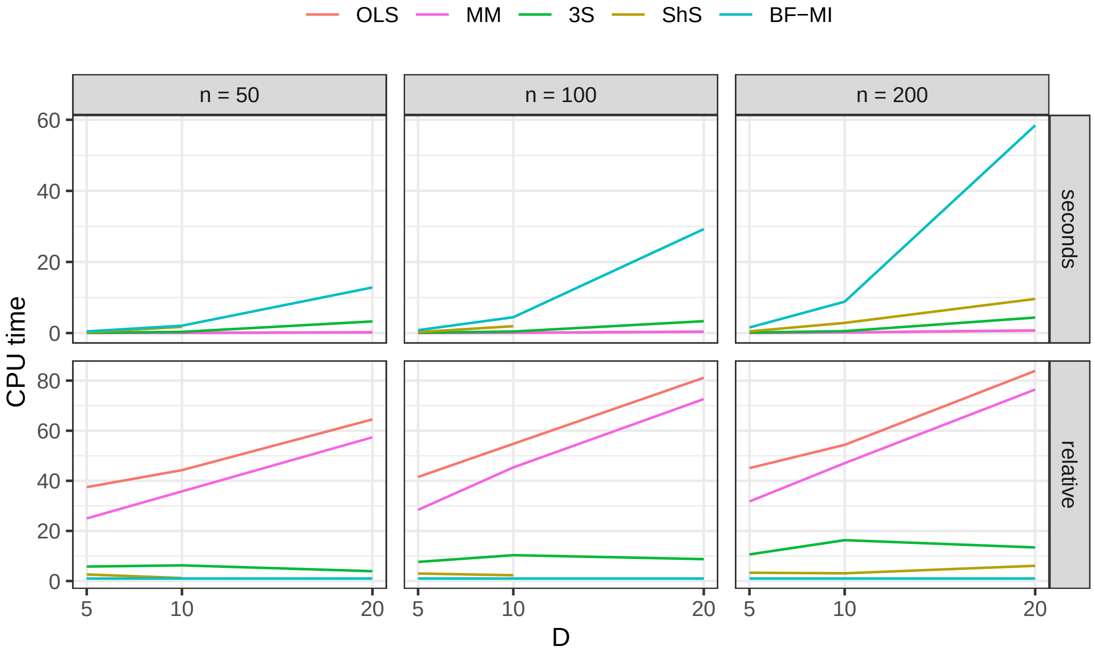

Fig. 3: Computation time of different methods for varying numbers of observations n and compositional parts D (averaged over 100 simulated data sets). The top row shows the computation time in seconds, while the bottom row shows the relative speed gain of each method with respect to our proposed method.

practical applications, in particular given that we do not consider the case of high-dimensional compositions. In our example with the VFA data ($n = 239$, $D = 6$, $p = 3$ real-valued covariates, 1 dummy variable), the computation time was 4.003 seconds; whereas compositional MM-regression required 0.113 seconds. It should also be noted that our current implementation is using R. It is likely that a considerable gain in speed can be achieved by implementing certain parts in, for example, the C++ language.

7 Conclusions and discussion

In compositional data analysis, the parts of a composition are considered intrinsically related to each other and the ratios between them constitute the key source of relevant information. However, cellwise outliers may be present in individual compositional parts. Keeping this problem in mind, we introduce a procedure to deal with cellwise outliers with the purpose of conducting robust regression analysis while taking the nature of compositional data into account. For the detection of cellwise outliers, we apply a bivariate filter (Rousseeuw and Van den Bossche, 2018) at the level of pairwise logratios where the elemental information is contained. For the imputation of cellwise outliers, we adapt an existing imputation method for missing compositional data (Hron et al., 2010), which treats the problem indirectly via pivot coordinates. Alternative missing data imputation methods could be developed in future research, e.g., robust versions of the non-parametric and Bayesian approaches implemented in the R package `zCompositions` (Palarea-

Albaladejo and Martín-Fernández, 2015). Importantly, using rowwise robust imputation and regression after filtering cellwise outliers yields a procedure that protects against both cellwise and rowwise outliers.

In our simulation study, the proposed BF-MI algorithm outperforms the well-known rowwise robust MM-estimator (Yohai, 1987) and the more recently introduced cellwise robust shooting S-estimator (Öllerer et al., 2016). In most simulation scenarios, the prediction performance of our method is similar to that of 3-step regression (Leung et al., 2016), which is another recent cellwise and rowwise robust regression proposal. Nevertheless, 3-step regression can only be applied to additive logratio (alr) coordinates, since pivot coordinates would turn cellwise outliers in the original data into entire outlying rows, which could easily render the majority of rows to be outliers. Even with alr coordinates, this outlier propagation occurs for observations with an outlying cell in the reference part in the denominator of the alr coordinates, but not for outlying cells in other parts. Moreover, applying 3-step regression to different alr-coordinate representations yields different predictions. The imputation stage in our method, going back to the original compositional parts, allows for predicted values that do not depend on a specific coordinate representation. In addition, the regression analysis can be done on any coordinate system that gives the desired interpretation. These advantages make our method preferable for practical purposes. Here we used pivot coordinates, which are particularly popular in the context of exploratory data analysis (Filzmoser et al., 2018), but other choices are possible, e.g., other orthonormal coordinates such as balances (Egozcue and Pawlowsky-Glahn, 2005) or weighted pivot coordinates (Hron et al., 2017), or even oblique coordinate systems (Greenacre, 2018).

Finally, some limitations of our proposed method are discussed. The simulation results for a (hypothetical) variation of the procedure with an ideal outlier filter are an indication that further refinement of the outlier filter could yield an improvement in performance. Hence this could be a fruitful venue for future research. Moreover, the procedure tends to become unstable when the number of variables approaches the number of observations, and it cannot be used when the number of variables is larger than the number of observations. For the latter case, estimators that are affine equivariant and rowwise robust are not available. This poses a challenge for high-dimensional compositional data and their coordinate representations, which are mutually related through affine transformations (rotations in case of orthonormal coordinates). Hence, the properties of the regression coefficient estimates after rotations of pivot coordinate systems (as shown in Section 3.3) are in general not satisfied, and alternative approaches would be needed. Thus, an extension of the proposed procedure to the high-dimensional case is a challenge for future research.

Acknowledgements This is a post-peer-review, pre-copyedit version of an article published in *Advances in Data Analysis and Classification*. The final authenticated version is available online at `https://doi.org/10.1007/s11634-021-00436-9`. We thank the editor and two anonymous referees for their valuable comments. N. Š. was supported by the Palacký University Grant Agency (IGA_PrF_2018_024 and IGA_PrF_2020_015) and the Scottish Government's Rural and Environment Science and Analytical Services Division. A. A. was supported by a grant of the Dutch Research Council (NWO), research program Vidi (project number VI.Vidi.195.141). J. P.-A. was supported by the Scottish Government's Rural and Environment Science and Analytical Services Division and the Spanish Ministry of Economy and Competitiveness (Ref: RTI2018-095518-B-C21). K. H. was supported by the Palacký University Grant Agency (IGA_PrF_2018_024 and IGA_PrF_2020_015) and by the Grant Agency of the Czech Republic (19-07155S).

References

Agostinelli C, Leung A, Yohai V, Zamar R (2015) Robust estimation of multivariate location and scatter in the presence of cellwise and casewise contamination. TEST 24(3):441–461

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

Allison P (2002) Missing Data. SAGE, Thousand Oaks

Alqallaf F, Van Aelst S, Yohai V, Zamar R (2009) Propagation of outliers in multivariate data. The Annals of Statistics 37(1):311–331

Barnard J, Rubin D (1999) Small-sample degrees of freedom with multiple imputation. Biometrika 86(4):948–955

Bodner T (2009) What improves with increased missing data imputations? Structural Equation Modeling: A Multidisciplinary Journal 15(4):651–675

Cevallos Valdiviezo H, Van Aelst S (2015) Tree-based prediction on incomplete data using imputation or surrogate decisions. Information Sciences 311:163–181

R Core Team (2020) R: A Language and Environment for Statistical Computing. R Foundation for Statistical Computing, Vienna, Austria, URL https://www.R-project.org

Danilov M, Yohai V, Zamar R (2012) Robust estimation of multivariate location and scatter in the presence of missing data. Journal of the American Statistical Association 107(499):1178–1186

Egozcue J, Pawlowsky-Glahn V (2005) Groups of parts and their balances in compositional data analysis. Mathematical Geology 37(7):795–828

Egozcue J, Pawlosky-Glahn V, Mateu-Figueras G, Barceló-Vidal C (2003) Isometric logratio transformations for compositional data analysis. Mathematical Geology 35(3):279–300

Farcomeni A (2014a) Robust constrained clustering in presence of entry-wise outliers. Technometrics 56(1):102–111

Farcomeni A (2014b) Snipping for robust k-means clustering under component-wise contamination. Statistics and Computing 24(6):907–919

Filzmoser P, Hron K, Templ M (2018) Applied Compositional Data Analysis. Springer, Cham

Filzmoser P, Höppner S, Ortner I, Serneels S, Verdonck T (2020) Cellwise robust M regression. Computational Statistics & Data Analysis 147:106944

Fišerová E, Hron K (2011) On the interpretation of orthonormal coordinates for compositional data. Mathematical Geosciences 43(4):455–468

Greenacre M (2018) Compositional Data Analysis in Practice. CRC Press, Boca Raton

Hampel F, Ronchetti E, Rousseeuw P, Stahel W (1986) Robust Statistics: The Approach Based on Influence Functions. John Wiley & Sons, New York

Hron K, Filzmoser P (2010) Elements of robust regression for data with absolute and relative information. In: Borgelt C, González-Rodríguez G, Trutschnig W, Lubiano M, Gil M, Grzegorzewski P, Hryniewicz O (eds) Combining Soft Computing and Statistical Methods in Data Analysis, Springer, Heidelberg, pp 329–335

Hron K, Templ M, Filzmoser P (2010) Imputation of missing values for compositional data using classical and robust methods. Computational Statistics & Data Analysis 54(12):3095–3107

Hron K, Filzmoser P, Thompson K (2012) Linear regression with compositional explanatory variables. Journal of Applied Statistics 39(5):1115–1128

Hron K, Filzmoser P, de Caritat P, Fišerová E, Gardlo A (2017) Weighted pivot coordinates for compositional data and their application to geochemical mapping. Mathematical Geosciences 49(6):797–814

Hrůzová K, Todorov V, Hron K, Filzmoser P (2016) Classical and robust orthogonal regression between parts of compositional data. Statistics: A Journal of Theoretical and Applied Statistics 50(6):1261–1275

Huber P, Ronchetti E (2009) Robust Statistics, 2nd edn. John Wiley & Sons, Hoboken

Hubert M, Rousseeuw P, Van den Bossche W (2019) MacroPCA: An all-in-one PCA method allowing for missing values as well as cellwise and rowwise outliers. Technometrics In print

Khan J, Van Aelst S, Zamar R (2007) Robust linear model selection based on least angle regression. Journal of the American Statistical Association 102(480):1289–1299

Leung A, Zhang H, Zamar R (2015) `robreg3S`: Three-Step Regression and Inference for Cellwise and Casewise Contamination. URL `https://CRAN.R-project.org/package=robreg3S`, R package version 0.3

Leung A, Zhang H, Zamar R (2016) Robust regression estimation and inference in the presence of cellwise and casewise contamination. Computational Statistics & Data Analysis 99:1–11

Leung A, Yohai V, Zamar R (2017) Multivariate location and scatter matrix estimation under cellwise and casewise contamination. Computational Statistics & Data Analysis 111:59–76

Little R (1992) Regression with missing X's: A review. Journal of the American Statistical Association 87(420):1227–1237

Little R, Rubin D (2002) Statistical Analysis with Missing Data, 2nd edn. John Wiley & Sons, Chichester

Lopuhaä H, Rousseeuw P (1991) Breakdown points of affine equivariant estimators of multivariate location and covariance matrices. The Annals of Statistics 19(1):229–248

Maronna R, Martin R, Yohai V (2002) Robust Statistics: Theory and Methods. John Wiley & Sons, Chichester

Mersmann O (2019) `microbenchmark`: Accurate Timing Functions. URL `https://CRAN.R-project.org/package=microbenchmark`, R package version 1.4-7

Müller I, Hron K, Fišerová E, Šmahaj J, Cakirpaloglu P, Vančáková J (2018) Interpretation of compositional regression with application to time budget analysis. Austrian Journal of Statistics 47(2):3–19

Öllerer V, Alfons A, Croux C (2016) The shooting S-estimator for robust regression. Computational Statistics 31(3):829–844

Palarea-Albaladejo J, Martín-Fernández J (2015) zCompositions — R package for multivariate imputation of left-censored data under a compositional approach. Chemometrics and Intelligent Laboratory Systems 143:85–96

Palarea-Albaladejo J, Rooke JA, Nevison IM, Dewhurst RJ (2017) Compositional mixed modeling of methane emissions and ruminal volatile fatty acids from individual cattle and multiple experiments. Journal of Animal Science 95(6):2467–2480

Pawlowsky-Glahn V, Egozcue J, Tolosana-Delgado R (2015) Modeling and analysis of compositional data. John Wiley & Sons, Chichester

Raymaekers J, Rousseeuw P, Van den Bossche W (2019) `cellWise`: Analyzing Data with Cellwise Outliers. URL `https://CRAN.R-project.org/package=cellWise`, R package version 2.1.0

Rousseeuw P, Van den Bossche W (2018) Detecting deviating data cells. Technometrics 60(2):135–145

Rousseeuw P, Leroy A (1987) Robust Regression and Outlier Detection. John Wiley & Sons, New York

Rubin D (1987) Multiple Imputation for Nonresponse in Surveys. John Wiley & Sons

Rubin D, Schenker M (1986) Multiple imputation for interval estimation from simple random samples with ignorable nonresponse. Journal of the American Statistical Association 81(394):366–374

Templ M, Hron K, Filzmoser P (2011a) `robCompositions`: An R-package for robust statistical analysis of compositional data. In: Pawlowsky-Glahn V, Buccianti A (eds) Compositional Data Analysis: Theory and Applications, John Wiley & Sons, pp 341–355

Templ M, Kowarik A, Filzmoser P (2011b) Iterative stepwise regression imputation using standard and robust methods. Computational Statistics & Data Analysis 55(10):2793–2806

Van Aelst S, Vandervieren E, Willems G (2011) Stahel-Donoho estimators with cellwise weights. Journal of Statistical Computation and Simulation 81(1):1–27

Van Buuren S (2012) Flexible Imputation of Missing Data. Chapman & Hall/CRC, Boca Raton

White I, Royston P, Wood A (2011) Multiple imputation using chained equations: Issues and guidance for practice. Statistics in Medicine 30(4):377–399

Wolin M (1960) A theoretical rumen fermentation balance. Journal of Dairy Science 43:1452–1459

Yohai V (1987) High breakdown point and high efficiency robust estimates for regression. The Annals of Statistics 15(2):642–656

A Pseudocode of the BF-MI algorithm

The pseudocode uses the same notation as defined in Section 3. Elements of a matrix or vector are indicated by subscripts, e.g., x_{ij} denotes the i-th element of vector $\boldsymbol{x}_j$. Furthermore, let $I_A(.)$ be the indicator function for a set A. Algorithm 1 describes rowwise robust compositional MM-regression, where it is important to keep in mind that MM-estimator can be seen as a weighted least squares estimator with data-dependent weights (Yohai, 1987). Algorithm 2 outlines cellwise outlier detection for compositional data, whereas Algorithms 3 and 4 describe the initial k-nearest-neighbor ($k$nn) imputation and the robust model-based imputation procedure, respectively. Finally, Algorithm 5 puts all the building blocks together for our proposed BF-MI procedure.

For simplicity, Algorithm 4 does not include the variable screening step for the imputation models (see Section 3.2). In addition, the output of Algorithm 5 is limited to the estimates of the interpretable regression coefficients (cf. Section 3.3) and the corresponding variance estimates. Significance tests for those coefficients can then be performed in the usual way for multiple imputation (see Barnard and Rubin, 1999). If one is interested in prediction, the algorithm can easily be adjusted in the following way. As all pivot coordinate systems yield the same predictions, it suffices to pick one set of pivot coordinates. One can then perform MM-regression with those pivot coordinates for each imputed data set, and average the coefficient estimates. For a new observation, the same pivot coordinates can be computed to obtain the prediction of the response with the averaged coefficients.

Algorithm 1 Compositional MM-regression

```
Input: Compositional data $\boldsymbol{X} = (\boldsymbol{x}_1, \dots, \boldsymbol{x}_D)$, real-valued covariates $\boldsymbol{V} = (\boldsymbol{v}_1, \dots, \boldsymbol{v}_p)$, real-valued response $\boldsymbol{y}$
Output: Regression coefficient estimates and corresponding variance estimates
▷ On the first pivot coordinate system, fully iterate the MM-regression algorithm
Compute pivot coordinates $\boldsymbol{z}_1^{(1)}, \dots, \boldsymbol{z}_{D-1}^{(1)}$ from $\boldsymbol{x}_1, \dots, \boldsymbol{x}_D$
Perform MM-regression of $\boldsymbol{y}$ on $\boldsymbol{z}_1^{(1)}, \dots, \boldsymbol{z}_{D-1}^{(1)}, \boldsymbol{v}_1, \dots, \boldsymbol{v}_p$
Store intercept $\hat{\alpha}$ and coefficient estimates $\hat{\beta}_1^{(1)}, \hat{\gamma}_1, \dots, \hat{\gamma}_p$ of variables $\boldsymbol{z}_1^{(1)}, \boldsymbol{v}_1, \dots, \boldsymbol{v}_p$, respectively
Compute variance estimates $\widehat{\mathrm{var}}(\hat{\alpha}), \widehat{\mathrm{var}}\left(\hat{\beta}_1^{(1)}\right), \widehat{\mathrm{var}}(\hat{\gamma}_1), \dots, \widehat{\mathrm{var}}(\hat{\gamma}_p)$
▷ The other pivot coordinate systems can use a weighted least squares fit
Obtain weights $\boldsymbol{w} = (w_1, \dots, w_n)'$ of observations from MM-regression fit
for $j \in \{2, \dots, D\}$ do
    Compute pivot coordinates $\boldsymbol{z}_1^{(j)}, \dots, \boldsymbol{z}_{D-1}^{(j)}$ from $\boldsymbol{x}_1, \dots, \boldsymbol{x}_D$
    Perform weighted least squares regression of $\boldsymbol{y}$ on $\boldsymbol{z}_1^{(j)}, \dots, \boldsymbol{z}_{D-1}^{(j)}, \boldsymbol{v}_1, \dots, \boldsymbol{v}_p$ with weights $\boldsymbol{w}$
    Store coefficient estimate $\hat{\beta}_1^{(j)}$ of coordinate $\boldsymbol{z}_1^{(j)}$
    Compute variance estimates $\widehat{\mathrm{var}}\left(\hat{\beta}_1^{(j)}\right)$
end for
▷ Return coefficient estimates and corresponding variance estimates
return $(\hat{\alpha}, \hat{\beta}_1^{(1)}, \dots, \hat{\beta}_1^{(D)}, \hat{\gamma}_1, \dots, \hat{\gamma}_p)'$ and $\left(\widehat{\mathrm{var}}(\hat{\alpha}), \widehat{\mathrm{var}}\left(\hat{\beta}_1^{(1)}\right), \dots, \widehat{\mathrm{var}}\left(\hat{\beta}_1^{(D)}\right), \widehat{\mathrm{var}}(\hat{\gamma}_1), \dots, \widehat{\mathrm{var}}(\hat{\gamma}_p)\right)'$
```

B On using a separate imputation step

As an alternative to the use of a separate imputation step, we looked into modifying the definition of pivot coordinates in (1) so that they account for missing values in the compositional data set. The main idea behind

Algorithm 2 Detection of cellwise outliers

```
Input: Data matrix $\boldsymbol{\mathcal{X}} = (\boldsymbol{x}_1, \dots, \boldsymbol{x}_D, \boldsymbol{r}_1, \dots, \boldsymbol{r}_{p+1})$ of compositional parts and real-valued variables
Output: Index set $\mathcal{O}$ of outlying cells and index set $\mathcal{U}$ of outlying rows
▷ Cellwise outlier detection on pairwise logratios and real-valued variables
$\boldsymbol{\mathcal{L}} \leftarrow (\ln(\boldsymbol{x}_1/\boldsymbol{x}_2), \dots, \ln(\boldsymbol{x}_{D-1}/\boldsymbol{x}_D), \boldsymbol{r}_1, \dots, \boldsymbol{r}_{p+1})$
Apply bivariate filter of Rousseeuw and Van den Bossche (2018) to $\boldsymbol{\mathcal{L}}$
Store index set $\mathcal{O}_{\boldsymbol{\mathcal{L}}} \leftarrow \{(i,k) :$ cell in row $i$ and column $k$ of $\boldsymbol{\mathcal{L}}$ is marked as cellwise outlier$\}$
Store index set $\mathcal{U}_{\boldsymbol{\mathcal{L}}} \leftarrow \{i :$ row $i$ of $\boldsymbol{\mathcal{L}}$ is marked as rowwise outlier$\}$
▷ Mark outlying cells in compositional parts
Initialize empty set $\mathcal{O}$          ▷ set of indices $(i,j)$ of cells in $\boldsymbol{\mathcal{X}}$ to be marked as cellwise outliers
Initialize empty set $\mathcal{U}$          ▷ set of indices $i$ of rows in $\boldsymbol{\mathcal{X}}$ to be marked as rowwise outliers
for $j \in \{1, \dots, D\}$ do
    Obtain index set $K_j \leftarrow \{k :$ column $k$ of $\boldsymbol{\mathcal{L}}$ contains a logratio involving $x_j\}$
    for $i \in \{1, \dots, n\}$ do
        if $\frac{1}{(D-1)} \sum_{k \in K_j} I_{\mathcal{O}_{\boldsymbol{\mathcal{L}}}}((i,k)) \geq 0.5$ then
            $\mathcal{O} \leftarrow \mathcal{O} \cup \{(i,j)\}$
        end if
    end for
end for
▷ Adopt outlying cells in real-valued variables from bivariate filter
for $j \in \{1, \dots, p+1\}$ do
    for $i \in \{1, \dots, n\}$ do
        if $(i, D(D-1)/2 + j) \in \mathcal{O}_{\boldsymbol{\mathcal{L}}}$ then
            $\mathcal{O} \leftarrow \mathcal{O} \cup \{(i, D+j)\}$
        end if
    end for
end for
▷ Mark outlying rows and only mark outlying cells that are not part of outlying rows
for $i \in \{1, \dots, n\}$ do
    if $i \in \mathcal{U}_{\boldsymbol{\mathcal{L}}}$ or $\frac{1}{D+p+1} \sum_{j=1}^{D+p+1} I_{\mathcal{O}}((i,j)) >= 0.75$ then
        ▷ Marked as rowwise outlier in $\boldsymbol{\mathcal{L}}$ or at least 75% of cells marked as cellwise outliers in $\boldsymbol{\mathcal{X}}$
        $\mathcal{U} \leftarrow \mathcal{U} \cup \{i\}$
        $\mathcal{O} \leftarrow \mathcal{O} \setminus \{(i,j) : j = 1, \dots, D+p+1\}$
    end if
end for
return Index sets $\mathcal{O}$ and $\mathcal{U}$
```

Algorithm 3 Initial $k$nn imputation for compositional data and real-valued variables

```
Input: Data matrix $\boldsymbol{\mathcal{X}} = (\boldsymbol{x}_1, \dots, \boldsymbol{x}_D, \boldsymbol{r}_1, \dots, \boldsymbol{r}_{p+1})$ of compositional parts and real-valued variables with missing values (outlying cells)
Output: Imputed data matrix $\tilde{\boldsymbol{\mathcal{X}}}$
Apply simultaneous $k$nn imputation with Aitchison distance to $\boldsymbol{X} = (\boldsymbol{x}_1, \dots, \boldsymbol{x}_D)$
Store imputed data matrix as $\tilde{\boldsymbol{X}} = (\tilde{\boldsymbol{x}}_1, \dots, \tilde{\boldsymbol{x}}_D)$
Compute pivot coordinates $\tilde{\boldsymbol{z}}_1^{(1)}, \dots, \tilde{\boldsymbol{z}}_{D-1}^{(1)}$ from $\tilde{\boldsymbol{x}}_1, \dots, \tilde{\boldsymbol{x}}_D$
Apply simultaneous $k$nn imputation with Euclidean distance to $\left(\boldsymbol{r}_1, \dots, \boldsymbol{r}_{p+1}, \tilde{\boldsymbol{z}}_1^{(1)}, \dots, \tilde{\boldsymbol{z}}_{D-1}^{(1)}\right)$
Store imputed real-valued variables as $\tilde{\boldsymbol{R}} = (\tilde{\boldsymbol{r}}_1, \dots, \tilde{\boldsymbol{r}}_{p+1})$
return Imputed data matrix $\tilde{\boldsymbol{\mathcal{X}}} = (\tilde{\boldsymbol{X}}, \tilde{\boldsymbol{R}})$
```

these missing value preserving pivot coordinates is that the geometric mean in the denominator of the logratio in (1) discards missing values.

Let $\mathbf{u}_i = (u_{i1}, \dots, u_{iD})'$ be an indicator vector for observed values in a composition $\mathbf{x}_i = (x_{i1}, \dots, x_{iD})'$, i.e.:

$$u_{ik} = \begin{cases} 1 & \text{if } x_{ik} \text{ is observed,} \\ 0 & \text{if } x_{ik} \text{ is missing,} \end{cases} \qquad k = 1, \dots, D.$$

Algorithm 4 Model-based imputation for compositional data and real-valued variables

```
Input: Data matrix $\boldsymbol{\mathcal{X}} = (\boldsymbol{x}_1, \dots, \boldsymbol{x}_D, \boldsymbol{r}_1, \dots, \boldsymbol{r}_{p+1})$ of compositional parts and real-valued variables with missing values (outlying cells)
Output: Imputed data matrix $\tilde{\boldsymbol{\mathcal{X}}}$, residual scale estimates $\hat{\sigma}_1, \dots, \hat{\sigma}_{D+p+1}$ from imputation models
▷ Initializations
Rearrange first $D$ columns of $\boldsymbol{\mathcal{X}}$ by sorting compositional parts by decreasing amount of missing values
Rearrange last $p+1$ columns of $\boldsymbol{\mathcal{X}}$ by sorting real-valued variables by decreasing amount of missing values
Obtain index sets $\phi_j \leftarrow \{i : \text{cell in row } i \text{ and column } j \text{ of } \boldsymbol{\mathcal{X}} \text{ is missing}\}$, $j = 1, \dots, D+p+1$
Obtain index sets $\psi_j \leftarrow \{i : \text{cell in row } i \text{ and column } j \text{ of } \boldsymbol{\mathcal{X}} \text{ is observed}\}$, $j = 1, \dots, D+p+1$
Initialize counter $it \leftarrow 0$ and convergence criterion $\eta \leftarrow \infty$
Initialize $\boldsymbol{\mathcal{X}}^{[0]} = \left(\boldsymbol{x}_1^{[0]}, \dots, \boldsymbol{x}_D^{[0]}, \boldsymbol{r}_1^{[0]}, \dots, \boldsymbol{r}_{p+1}^{[0]}\right)$ by applying $k$nn imputation from Algorithm 3 to $\boldsymbol{\mathcal{X}}$
▷ Iterative model-based imputations
while $\eta \geq 0.5$ do
    $it \leftarrow it + 1$
    $\boldsymbol{\mathcal{X}}^{[it]} = \left(\boldsymbol{x}_1^{[it]}, \dots, \boldsymbol{x}_D^{[it]}, \boldsymbol{r}_1^{[it]}, \dots, \boldsymbol{r}_{p+1}^{[it]}\right) \leftarrow \boldsymbol{\mathcal{X}}^{[it-1]} = \left(\boldsymbol{x}_1^{[it-1]}, \dots, \boldsymbol{x}_D^{[it-1]}, \boldsymbol{r}_1^{[it-1]}, \dots, \boldsymbol{r}_{p+1}^{[it-1]}\right)$
    ▷ Imputations in compositional data
    for $j \in \{1, \dots, D\}$ do
        Compute pivot coordinates $z_{i1}^{(j)}, \dots, z_{i,D-1}^{(j)}$ from $x_{i1}^{[it]}, \dots, x_{iD}^{[it]}$, $i = 1, \dots, n$
        Perform MM-regression of $z_{i1}^{(j)}$ on $z_{i2}^{(j)}, \dots, z_{i,D-1}^{(j)}, r_{i1}^{[it]}, \dots, r_{i,p+1}^{[it]}$, $i \in \psi_j$
        Compute prediction $\hat{z}_{i1}^{(j)}$ from $z_{i2}^{(j)}, \dots, z_{i,D-1}^{(j)}, r_{i1}^{[it]}, \dots, r_{i,p+1}^{[it]}$, $i \in \phi_j$
        Replace $x_{i1}^{[it]}, \dots, x_{iD}^{[it]}$ with the inverse mapping of $\hat{z}_{i1}^{(j)}, z_{i2}^{(j)}, \dots, z_{i,D-1}^{(j)}$, $i \in \phi_j$
        Compute robust residual scale estimate $\hat{\sigma}_j$ from MM-regression fit
    end for
    ▷ Imputations in real-valued variables
    Compute pivot coordinates $z_{i1}^{(1)}, \dots, z_{i,D-1}^{(1)}$ from $x_{i1}^{[it]}, \dots, x_{iD}^{[it]}$, $i = 1, \dots, n$
    for $j \in \{1, \dots, p+1\}$ do
        Perform MM-regression of $r_{ij}^{[it]}$ on $z_{i1}^{(1)}, \dots, z_{i,D-1}^{(1)}, r_{i1}^{[it]}, \dots, r_{i,j-1}^{[it]}, r_{i,j+1}^{[it]}, r_{i,p+1}^{[it]}$, $i \in \psi_j$
        Replace $r_{ij}^{[it]}$ with prediction $\hat{r}_{ij}^{[it]}$ from $z_{i1}^{(1)}, \dots, z_{i,D-1}^{(1)}, r_{i1}^{[it]}, \dots, r_{i,j-1}^{[it]}, r_{i,j+1}^{[it]}, r_{i,p+1}^{[it]}$, $i \in \phi_j$
        Compute robust residual scale estimate $\hat{\sigma}_{D+j}$ from MM-regression fit
    end for
    ▷ Update convergence criterion
    $\eta \leftarrow \sum_{i=1}^{n} \left[ \sum_{j=1}^{D} \left( \frac{x_{ij}^{[it-1]} - x_{ij}^{[it]}}{x_{ij}^{[it]}} \right)^2 + \sum_{j=1}^{p+1} \left( \frac{r_{ij}^{[it-1]} - r_{ij}^{[it]}}{r_{ij}^{[it]}} \right)^2 \right]$
end while
Obtain $\tilde{\boldsymbol{\mathcal{X}}}$ by rearranging columns of $\boldsymbol{\mathcal{X}}^{[it]}$ from last iteration according to original order of columns in $\boldsymbol{\mathcal{X}}$
Rearrange residual scale estimates $\hat{\sigma}_1, \dots, \hat{\sigma}_{D+p+1}$ accordingly
return Imputed data matrix $\tilde{\boldsymbol{\mathcal{X}}}$ and residual scale estimates $\hat{\sigma}_1, \dots, \hat{\sigma}_{D+p+1}$
```

Then $D(\mathbf{u}_i) = \sum_{l=1}^{D} u_{il}$ is the observed dimension of observation $\mathbf{x}_i$. With an indicator vector $\mathbf{v}_i = (v_{i1}, \dots, v_{i,D-1})'$ defined by

$$v_{ik} = \begin{cases} 1 & \text{if } u_{ik} = 1 \text{ and any } u_{il} = 1 \text{ for } l > k, \\ 0 & \text{otherwise,} \end{cases} \qquad k = 1, \dots, D-1,$$

we can define missing value preserving pivot coordinates $\tilde{\mathbf{z}}_i = (\tilde{z}_{i1}, \dots, \tilde{z}_{i,D-1})'$ as

$$\tilde{z}_{ik} = \begin{cases} \sqrt{\frac{D(\mathbf{u}_i) - j_k(\mathbf{v}_i)}{D(\mathbf{u}_i) - j_k(\mathbf{v}_i) + 1}} \ln \left(\frac{x_{ik}}{\sqrt[D(\mathbf{u}_i) - j_k(\mathbf{v}_i)]{\prod_{l>k:\, u_{il}=1} x_{il}}} \right) & \text{if } v_{ik} = 1, \\ \text{missing} & \text{if } v_{ik} = 0, \end{cases} \qquad k = 1, \dots, D-1, \tag{8}$$

where $j_k(\mathbf{v}_i) = \sum_{l=1}^{k} v_{il}$. Note that $j_k(\mathbf{v}_i)$ accounts for the number of observed pivot coordinates up to and including the current coordinate.

Now consider the regression model

$$Y = \alpha + \beta_1 Z_1 + \ldots + \beta_{D-1} Z_{D-1} + \varepsilon$$

Algorithm 5 Robust compositional regression with bivariate filter and multiple imputation

```
Input: Compositional data $\boldsymbol{X} = (\boldsymbol{x}_1, \dots, \boldsymbol{x}_D)$, real-valued covariates $\boldsymbol{V} = (\boldsymbol{v}_1, \dots, \boldsymbol{v}_p)$, real-valued response $\boldsymbol{y}$
Output: Regression coefficient estimates and corresponding variance estimates
▷ Detect cellwise outliers
Obtain index set $\mathcal{O}$ of cellwise outliers by applying Algorithm 2 to $\boldsymbol{\mathcal{X}} = (\boldsymbol{x}_1, \dots, \boldsymbol{x}_D, \boldsymbol{v}_1, \dots, \boldsymbol{v}_p, \boldsymbol{y})$
▷ Special case of no cellwise outliers
if $\mathcal{O} = \emptyset$ then
    Apply Algorithm 1 for compositional MM-regression of $\boldsymbol{y}$ on $\boldsymbol{x}_1, \dots, \boldsymbol{x}_D, \boldsymbol{v}_1, \dots, \boldsymbol{v}_p$
    return Coefficient estimates and corresponding variance estimates
end if
▷ Filter and impute cellwise outliers
Replace cells of $\boldsymbol{\mathcal{X}}$ with indices in $\mathcal{O}$ by missing values
Apply model-based imputation with Algorithm 4 to $\boldsymbol{\mathcal{X}} = (\boldsymbol{x}_1, \dots, \boldsymbol{x}_D, \boldsymbol{v}_1, \dots, \boldsymbol{v}_p, \boldsymbol{y})$
Store imputed data matrix as $\tilde{\boldsymbol{\mathcal{X}}} = (\tilde{\boldsymbol{x}}_1, \dots, \tilde{\boldsymbol{x}}_D, \tilde{\boldsymbol{v}}_1, \dots, \tilde{\boldsymbol{v}}_p, \tilde{\boldsymbol{y}})$
Store residual scale estimates from imputation models as $\hat{\sigma}_1, \dots, \hat{\sigma}_{D+p+1}$, respectively
▷ Robust compositional regression with multiple imputation
$n_{\text{out}} \leftarrow n - \sum_{i=1}^{n} \prod_{j=1}^{D+p+1} (1 - I_{\mathcal{O}}((i, j)))$          ▷ Number of observations with outlying cells
$M \leftarrow \max(2, \text{round}(100 \cdot n_{\text{out}}/n))$          ▷ Number of imputations
Obtain $o_j \leftarrow \sum_{i=1}^{n} I_{\mathcal{O}}((i, j))$, $j = 1, \dots, D + p + 1$          ▷ Number of outlying cells per variable
for $m \in \{1, \dots, M\}$ do
    ▷ Add random noise to imputations
    Initialize $\tilde{\boldsymbol{\mathcal{X}}}^{\{m\}} = \left(\tilde{\boldsymbol{x}}_1^{\{m\}}, \dots, \tilde{\boldsymbol{x}}_D^{\{m\}}, \tilde{\boldsymbol{v}}_1^{\{m\}}, \dots, \tilde{\boldsymbol{v}}_p^{\{m\}}, \tilde{\boldsymbol{y}}^{\{m\}}\right)$ by $\tilde{\boldsymbol{\mathcal{X}}} = (\tilde{\boldsymbol{x}}_1, \dots, \tilde{\boldsymbol{x}}_D, \tilde{\boldsymbol{v}}_1, \dots, \tilde{\boldsymbol{v}}_p, \tilde{\boldsymbol{y}})$
    for $(i, j) \in \mathcal{O}$ do
        Draw random noise term $e \sim N(0, \hat{\sigma}_j^2(1 + o_j/n))$
        if $j \in \{1, \dots, D\}$ then          ▷ Compositional parts
            Compute pivot coordinates $\tilde{z}_{i1}^{(j)}, \dots, \tilde{z}_{i,D-1}^{(j)}$ from $\tilde{x}_{i1}, \dots, \tilde{x}_{iD}$
            $\tilde{z}_{i1}^{(j)} \leftarrow \tilde{z}_{i1}^{(j)} + e$
            Replace $\tilde{x}_{i1}^{\{m\}}, \dots, \tilde{x}_{iD}^{\{m\}}$ with the inverse mapping of $\tilde{z}_{i1}^{(j)}, \dots, \tilde{z}_{i,D-1}^{(j)}$
        else if $j \in \{D + 1, \dots, D + p\}$ then          ▷ Real-valued variables
            $\tilde{v}_{i,j-D}^{\{m\}} \leftarrow \tilde{v}_{i,j-D} + e$
        else          ▷ Response variable
            $\tilde{y}_i^{\{m\}} \leftarrow \tilde{y}_i + e$
        end if
    end for
    ▷ Perform compositional MM-regression
    Apply Algorithm 1 for compositional MM-regression of $\tilde{\boldsymbol{y}}^{\{m\}}$ on $\tilde{\boldsymbol{x}}_1^{\{m\}}, \dots, \tilde{\boldsymbol{x}}_D^{\{m\}}, \tilde{\boldsymbol{v}}_1^{\{m\}}, \dots, \tilde{\boldsymbol{v}}_p^{\{m\}}$
    Store coefficient estimates as $\hat{\boldsymbol{\theta}}^{\{m\}} = \left(\hat{\theta}_0^{\{m\}}, \dots, \hat{\theta}_{D+p}^{\{m\}}\right)'$
    Store variance estimates as $\hat{\boldsymbol{U}}^{\{m\}} = \left(\hat{U}_0^{\{m\}}, \dots, \hat{U}_{D+p}^{\{m\}}\right)'$
end for
▷ Aggregate results from multiple imputation
Compute final coefficient estimates $\hat{\theta}_j \leftarrow \frac{1}{M} \sum_{m=1}^{M} \hat{\theta}_j^{\{m\}}$, $j = 0, \dots, D + p$
Compute average within-imputation variances $\hat{W}_j \leftarrow \frac{1}{M} \sum_{m=1}^{M} \hat{U}_j^{\{m\}}$, $j = 0, \dots, D + p$
Compute between-imputation variances $\hat{B}_j \leftarrow \frac{1}{M-1} \sum_{m=1}^{M} \left(\hat{\theta}_j^{\{m\}} - \hat{\theta}_j\right)^2$, $j = 0, \dots, D + p$
Compute variance estimates $\hat{V}_j \leftarrow \hat{W}_j + \frac{M+1}{M} \hat{B}_j$, $j = 0, \dots, D + p$
return Coefficient estimates $(\hat{\theta}_0, \dots, \hat{\theta}_{D+p})'$ and corresponding variance estimates $(\hat{V}_0, \dots, \hat{V}_{D+p})'$
```

with error term ε, and denote $\boldsymbol{\beta} = (\beta_1, \dots, \beta_{D-1})'$. For simplicity we do not consider additional real-valued covariates in this section. With the missing value preserving pivot coordinates from (8), we obtain a data set $\left(y_i, \tilde{\boldsymbol{z}}_i'\right)_{1 \leq i \leq n}$. We can first compute the sample mean omitting missing values by coordinate, denoted by $\boldsymbol{m}$, and the sample covariance matrix based on pairwise complete observations, denoted by $\boldsymbol{S}$. With

$$\boldsymbol{m} = \begin{pmatrix} m_Y \\ \boldsymbol{m}_Z \end{pmatrix} \qquad \text{and} \qquad \boldsymbol{S} = \begin{pmatrix} S_{YY} & \boldsymbol{S}_{YZ} \\ \boldsymbol{S}_{ZY} & \boldsymbol{S}_{ZZ} \end{pmatrix},$$

estimates of the regression coefficients can be computed as $\hat{\boldsymbol{\beta}} = \boldsymbol{S}_{ZZ}^{-1}\boldsymbol{S}_{ZY}$ and $\hat{\alpha} = m_Y - \boldsymbol{m}_Z'\hat{\boldsymbol{\beta}}$. If the missing values are missing completely at random (MCAR), $\boldsymbol{m}$ and $\boldsymbol{S}$ are consistent estimators (Little and Rubin, 2002, p. 42–43), and therefore we have consistency of $\hat{\alpha}$ and $\hat{\boldsymbol{\beta}}$ under the usual assumptions of the linear regression model. However, the pairwise complete sample covariance matrix $\boldsymbol{S}$ is not guaranteed to be positive definite, so an eigenvalue correction may be necessary. For robust estimates of $\boldsymbol{m}$ and $\boldsymbol{S}$, one could use the generalized S-estimator of Danilov et al. (2012), which is also consistent under MCAR.

For linear regression with compositional explanatory variables, we are interested in all regression models based on the different pivot coordinate systems from (3), i.e.,

$$Y = \alpha + \beta_1^{(l)} Z_1^{(l)} + \ldots + \beta_{D-1}^{(l)} Z_{D-1}^{(l)} + \varepsilon, \qquad l = 1, \ldots, D.$$

With $\mathbf{x}_i^{(l)} = \left(x_{i1}^{(l)}, \ldots, x_{iD}^{(l)}\right)' = (x_{il}, x_{i2}, \ldots, x_{i,l-1}, x_{i,l+1}, \ldots, x_{iD})'$, we obtain different sets of missing value preserving pivot coordinates $\tilde{\mathbf{z}}_i^{(l)} = \left(\tilde{z}_{i1}^{(l)}, \ldots, \tilde{z}_{i,D-1}^{(l)}\right)'$, $l = 1, \ldots, D$, analogous to (8). Then the regression estimator based on the pairwise complete sample covariance matrix or the robust generalized S-estimator yields consistent estimates $\hat{\alpha}^{(l)}, \hat{\beta}_1^{(l)}, \ldots, \hat{\beta}_{D-1}^{(l)}$ under MCAR.

However, consistency is not enough in the context of compositional data, we also need to ensure that the properties of pivot coordinates hold on finite samples. The most important property is that the different pivot coordinate systems from (3) are orthogonal rotations of each other. This property is crucial for the interpretation of regression coefficients, and it will ensure that the estimates of the intercept (and any coefficients of additional real-valued explanatory variables) are identical. Furthermore, it does not matter which coordinate system is used for prediction purposes, as the predictions will be identical.

We therefore ran a small simulation study. We generated $n = 250$ observations on $D = 6$ compositional parts. We first generate observations $\mathbf{z}_i$ on $D - 1$ coordinates from a multivariate normal distribution with mean 0 and covariance matrix $\Sigma = (0.5^{|i-j|})_{1 \leq i,j \leq D-1}$. Then we generated the response $y_i = \mathbf{z}_i'\boldsymbol{\beta} + \varepsilon_i$ with $\boldsymbol{\beta} = (1, 0, 1, 0, 1)'$ and standard normal error terms ε_i. Afterwards, we applied the inverse ilr mapping to the coordinates $\mathbf{z}_i$ to obtain the D-part compositions $\mathbf{x}_i$. Finally, we set cells in the data matrix $\boldsymbol{X} = (\mathbf{x}_1', \ldots, \mathbf{x}_n')'$ to missing values with a probability of 10% (MCAR). We repeat this 500 times.

For each generated data set, we computed two different missing value preserving coordinate systems: one where the first coordinate is based on the first compositional part and one where the first coordinate is based on the second compositional part. Then we estimated the regression coefficients based on the pairwise complete sample covariance matrix. We generated n independent test observations in the same way as described above, and computed predictions using the coefficient estimates based on the two pivot coordinate systems.

As argued above, those predictions need to be identical, otherwise the missing value preserving pivot coordinates are not useful in practice. However, we obtain an average absolute difference of the predictions of 0.0172 (averaged over all simulation runs and all observations in the test set).

While the difference is on average small, the predictions are definitely not identical. Therefore the coordinate systems are not exact orthogonal rotations of each other, and we cannot use such missing value preserving coordinates in practice. For interpretability purposes and to obtain identical predictions from different pivot coordinate systems, we first need to impute the compositional data and then compute the pivot coordinates based on the complete (imputed) data, as proposed in Section 3.

C On including the response variable in the bivariate filter and multiple imputation

One point of discussion brought up by an anonymous reviewer is whether the response variable should be included in the cellwise outlier filter and subsequent multiple imputation, or whether outliers in the response should remain in the data to be treated by the MM-regression estimator. As we use a bivariate filter to detect cellwise outliers, including the response in the bivariate filter can help with detecting outlying cells in the explanatory variables more accurately. For multiple imputation, Allison (2002, p.53) argues that the response variable needs to be included in the imputation process. The main argument is that if the response variable is omitted *a priori* from the imputation models, the conditional distribution from which the imputed values are drawn will be in general misspecified, resulting in bias in the regression coefficients.

To further investigate this issue, we perform simulations to assess the effect of the imputations on the estimation of the model. We use the same simulation design as in Section 4, but we only consider $D = 5$ compositional parts. We keep the probability of rowwise outliers fixed at 0.05, but we vary the probability of cellwise outliers from 0 to 0.2. As a baseline, we use the robust MM-regression estimator applied to pivot coordinates on the simulated data without cellwise contamination. To isolate the effect of multiple imputation on the coefficient estimates, we apply the hypothetical version of the proposed procedure with an ideal outlier filter (IF-MI) to the same simulated data but with added cellwise contamination. Moreover, we also apply a variant of the procedure

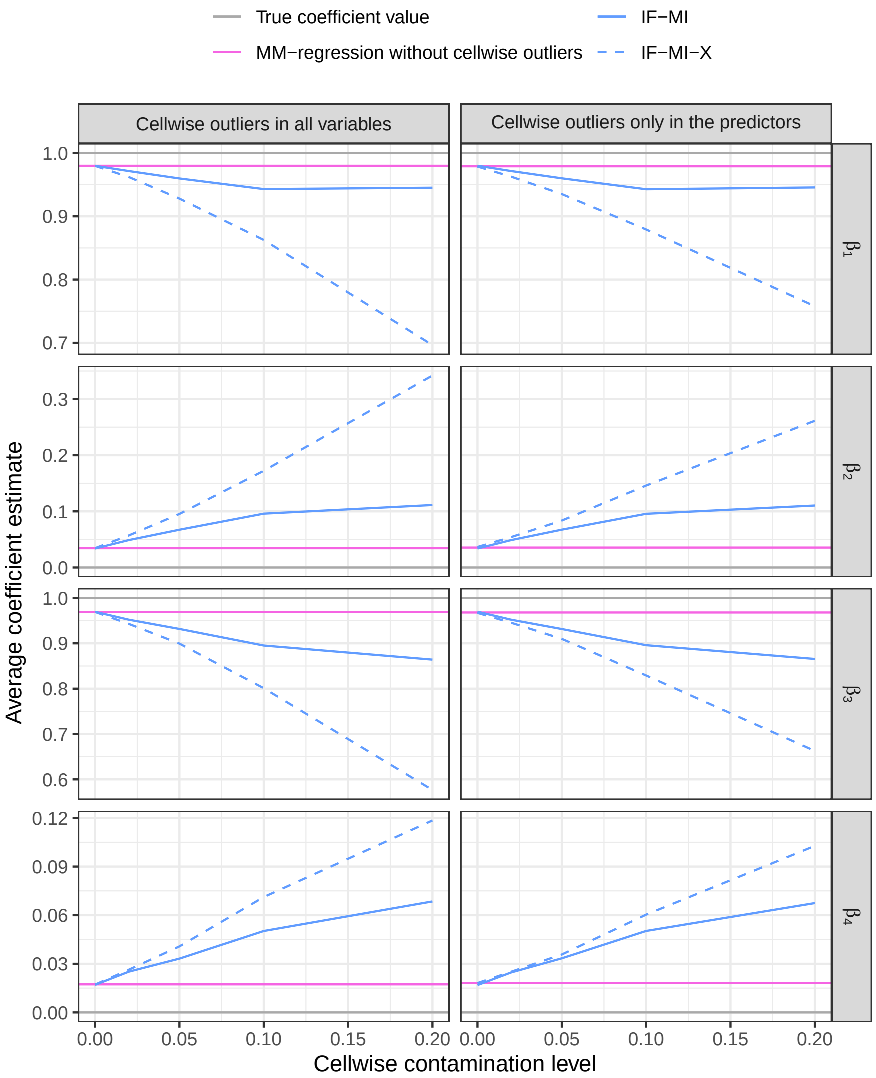

Fig. 4: Results from 1000 simulation runs to investigate the effect of the imputations on the estimated regression coefficients: the average coefficient estimates from regression methods in pivot coordinates are plotted against the cellwise contamination level for different cellwise outlier settings.

(IF-MI-X) that includes only the explanatory variables in the filtering and imputation steps, and leaves outliers in the response variable to be treated by the MM-estimator. In addition to generating cellwise outliers in all variables, we also consider a scenario with cellwise outliers only in the compositional explanatory variables. This second scenario isolates the role of the response variable as a predictor in the imputation models of IF-MI, since no values in the response are imputed.

Figure 4 shows the results for the $D - 1 = 4$ regression coefficients for sample size $n = 100$ and the multiplication factor of the covariance matrix $k = 1$. Results for other values of n and k lead to the same conclusions and are therefore omitted. By leaving out the correlated response variable in the multiple imputation process, bias in the regression coefficients is indeed amplified using IF-MI-X. As expected, this bias increases as the cellwise contamination level increases. On the other hand, when the response variable is included in the multiple imputations, the conditional distributions are more accurately modeled, and the bias of IF-MI is much lower.

D Figures of simulation results

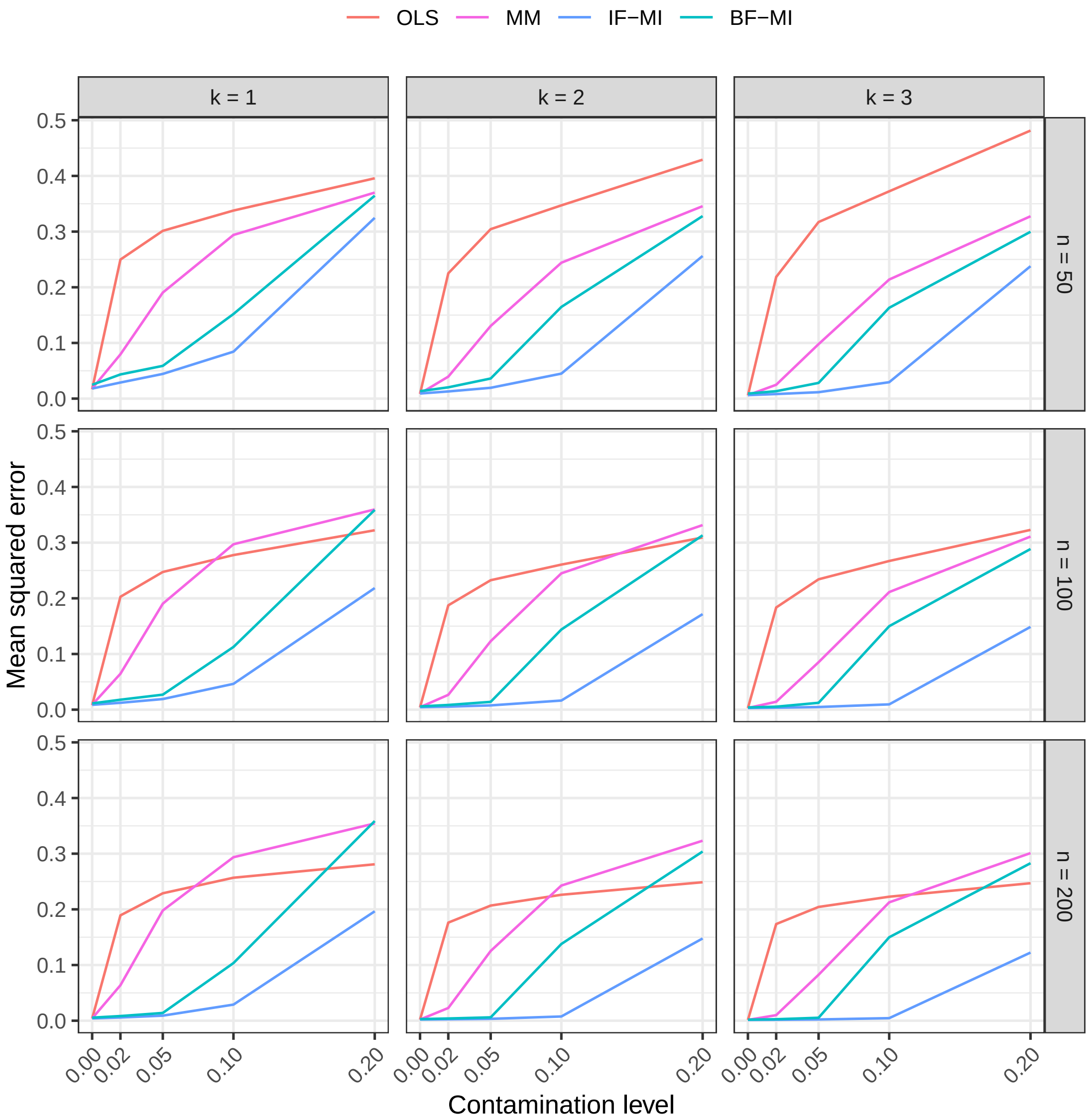

Fig. 5: Results from 1000 simulation runs for the scenario with $D = 5$ compositional parts: the average MSE of coefficient estimates from regression methods in pivot coordinates is plotted against the contamination level ζ for various sample sizes n and scaling factors k of the covariance matrix.

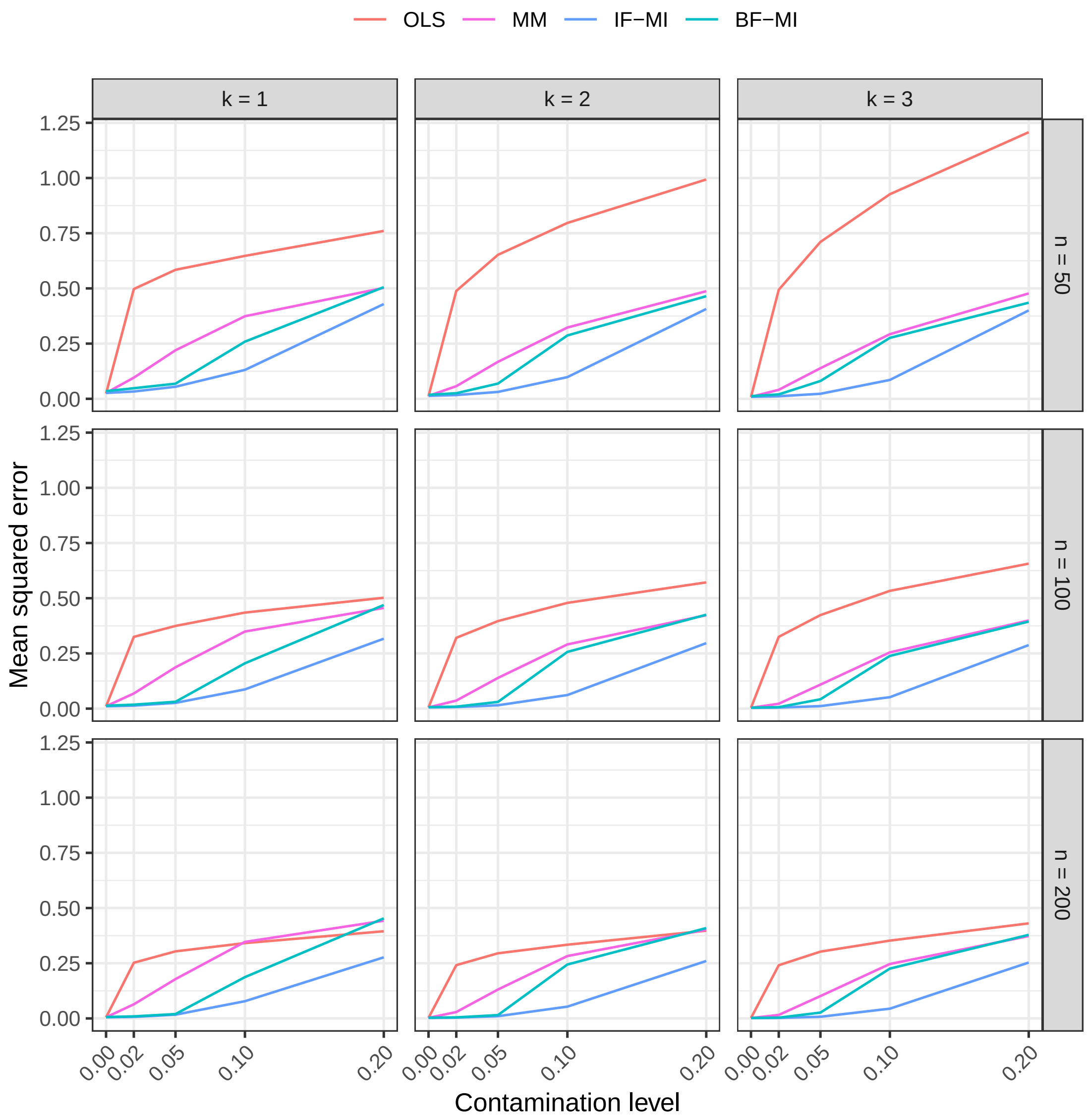

Fig. 6: Results from 1000 simulation runs for the scenario with $D = 10$ compositional parts: the average MSE of coefficient estimates from regression methods in pivot coordinates is plotted against the contamination level ζ for various sample sizes n and scaling factors k of the covariance matrix.

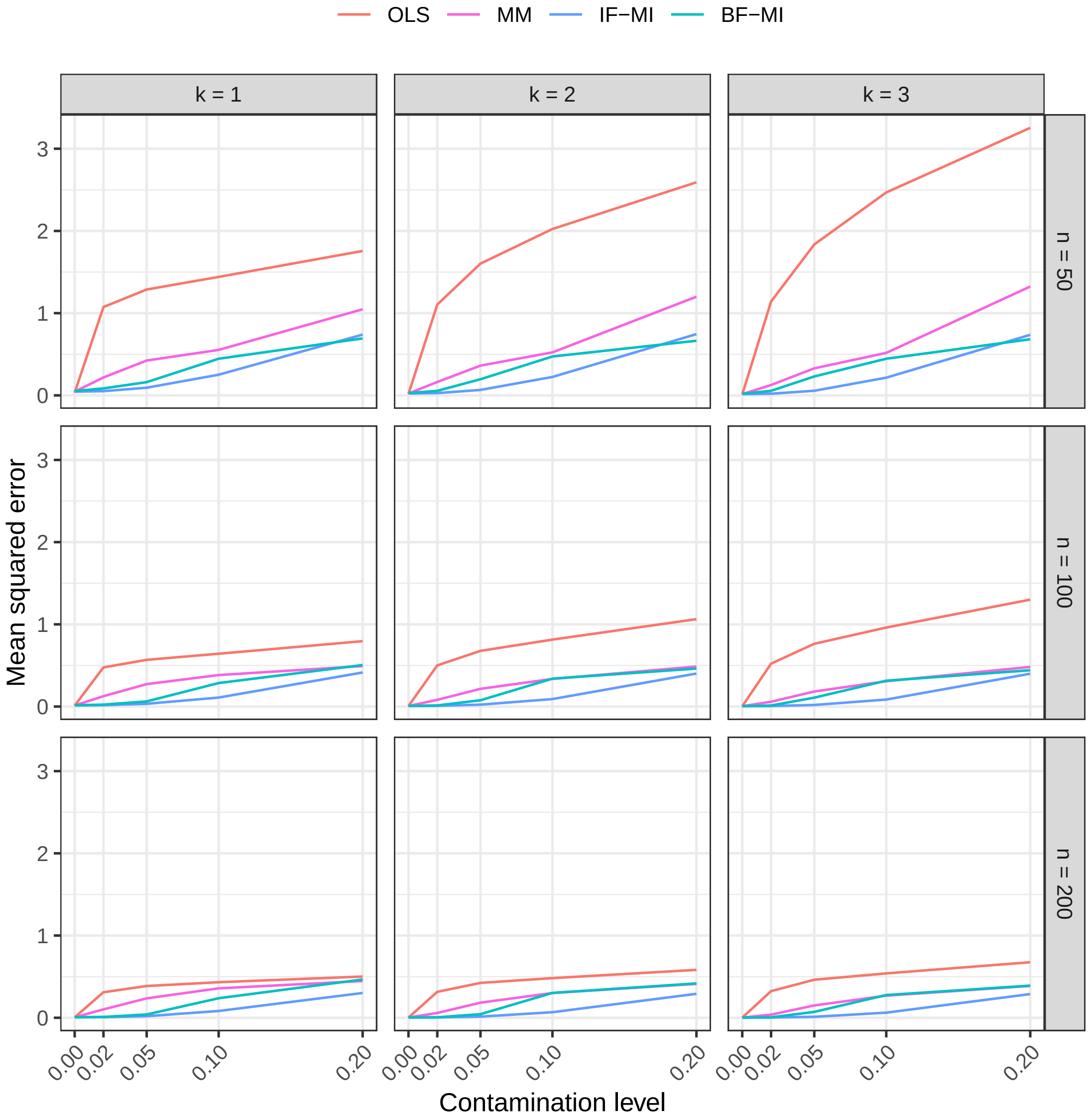

Fig. 7: Results from 1000 simulation runs for the scenario with $D = 20$ compositional parts: the average MSE of coefficient estimates from regression methods in pivot coordinates is plotted against the contamination level ζ for various sample sizes n and scaling factors k of the covariance matrix.

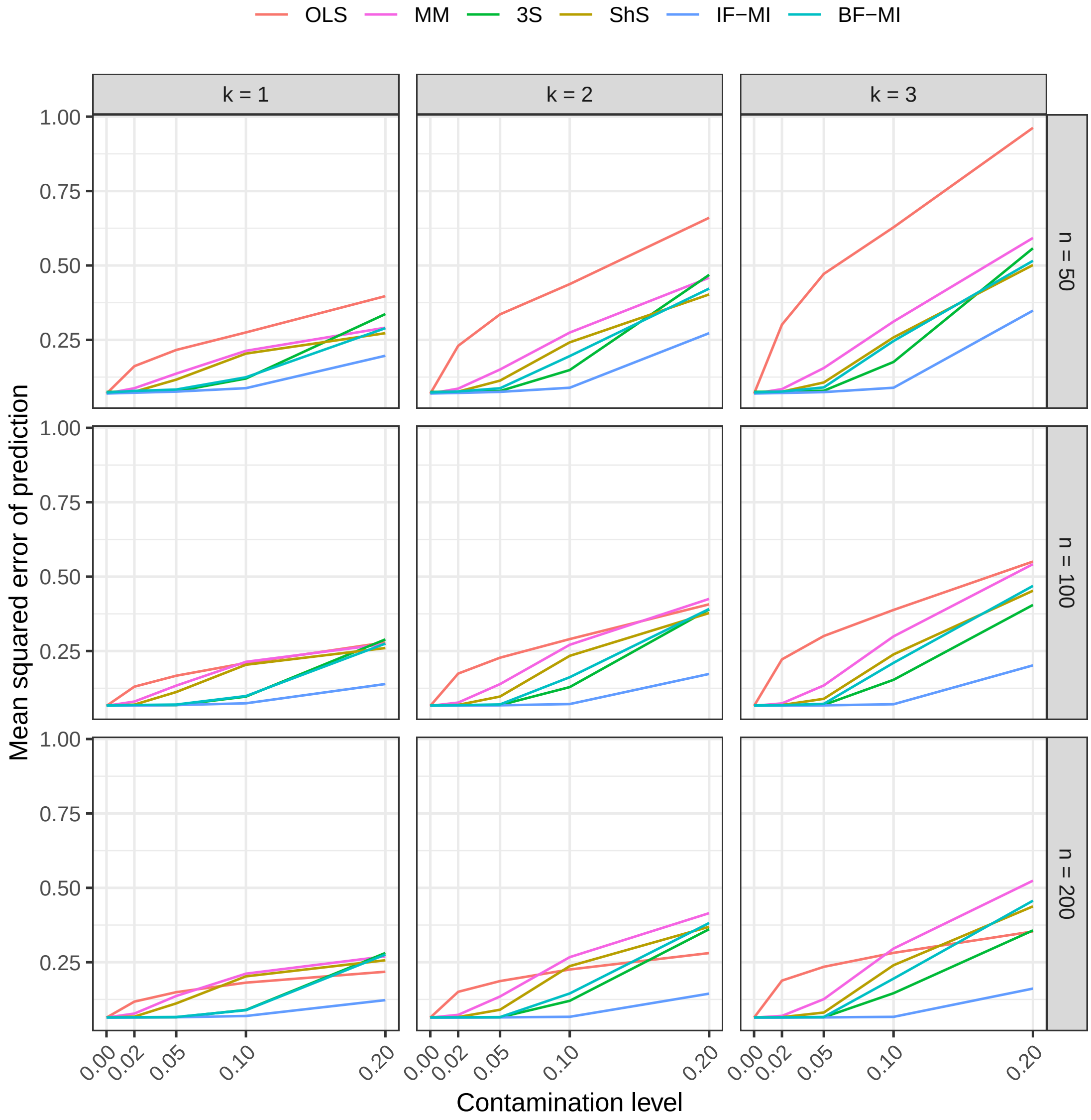

Fig. 8: Results from 1000 simulation runs for the scenario with $D = 5$ compositional parts: the average MSEP for different regression methods is plotted against the contamination level ζ for various sample sizes n and scaling factors k of the covariance matrix in pivot coordinates.

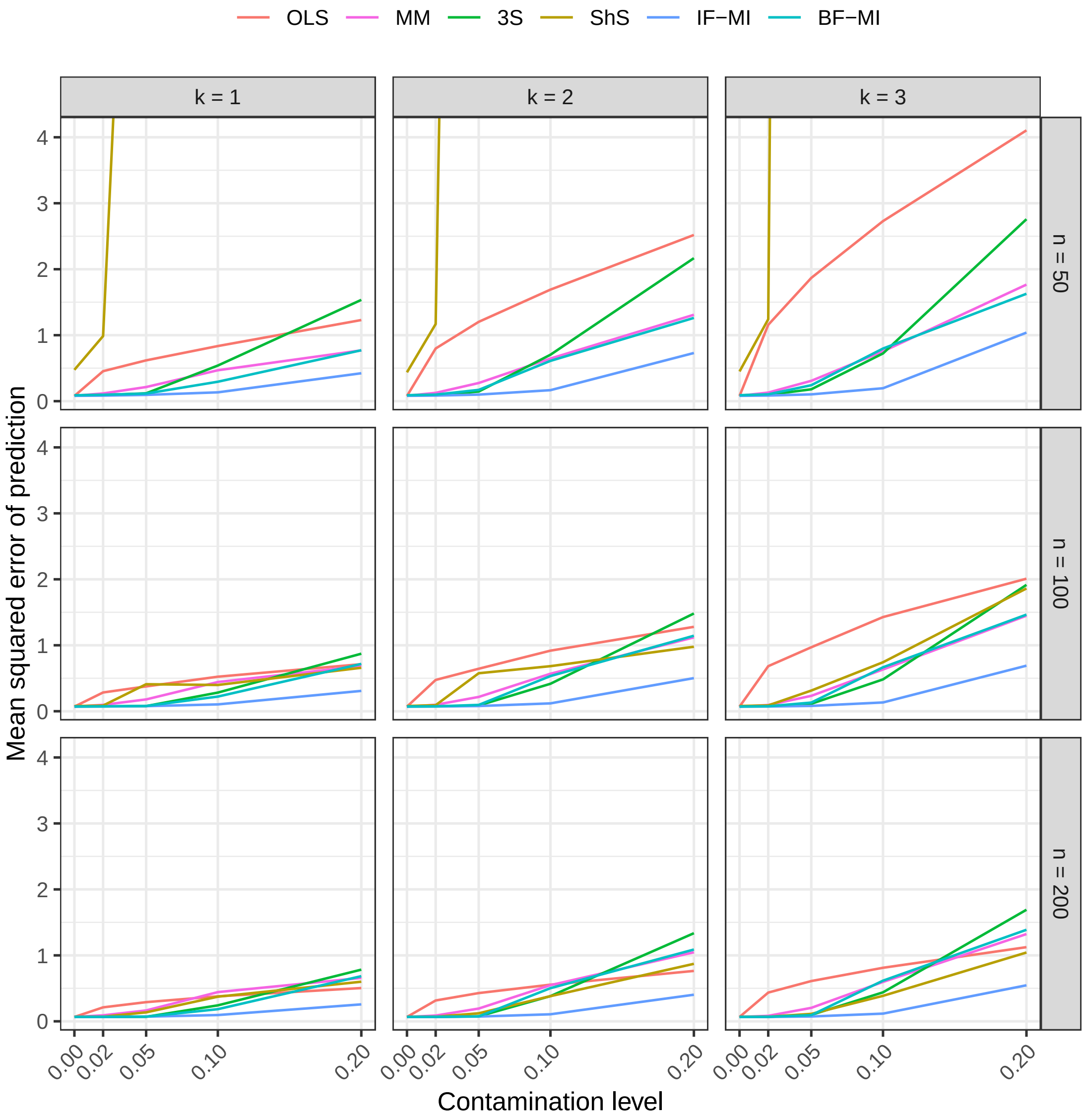

Fig. 9: Results from 1000 simulation runs for the scenario with $D = 10$ compositional parts: the average MSEP for different regression methods is plotted against the contamination level ζ for various sample sizes n and scaling factors k of the covariance matrix in pivot coordinates.

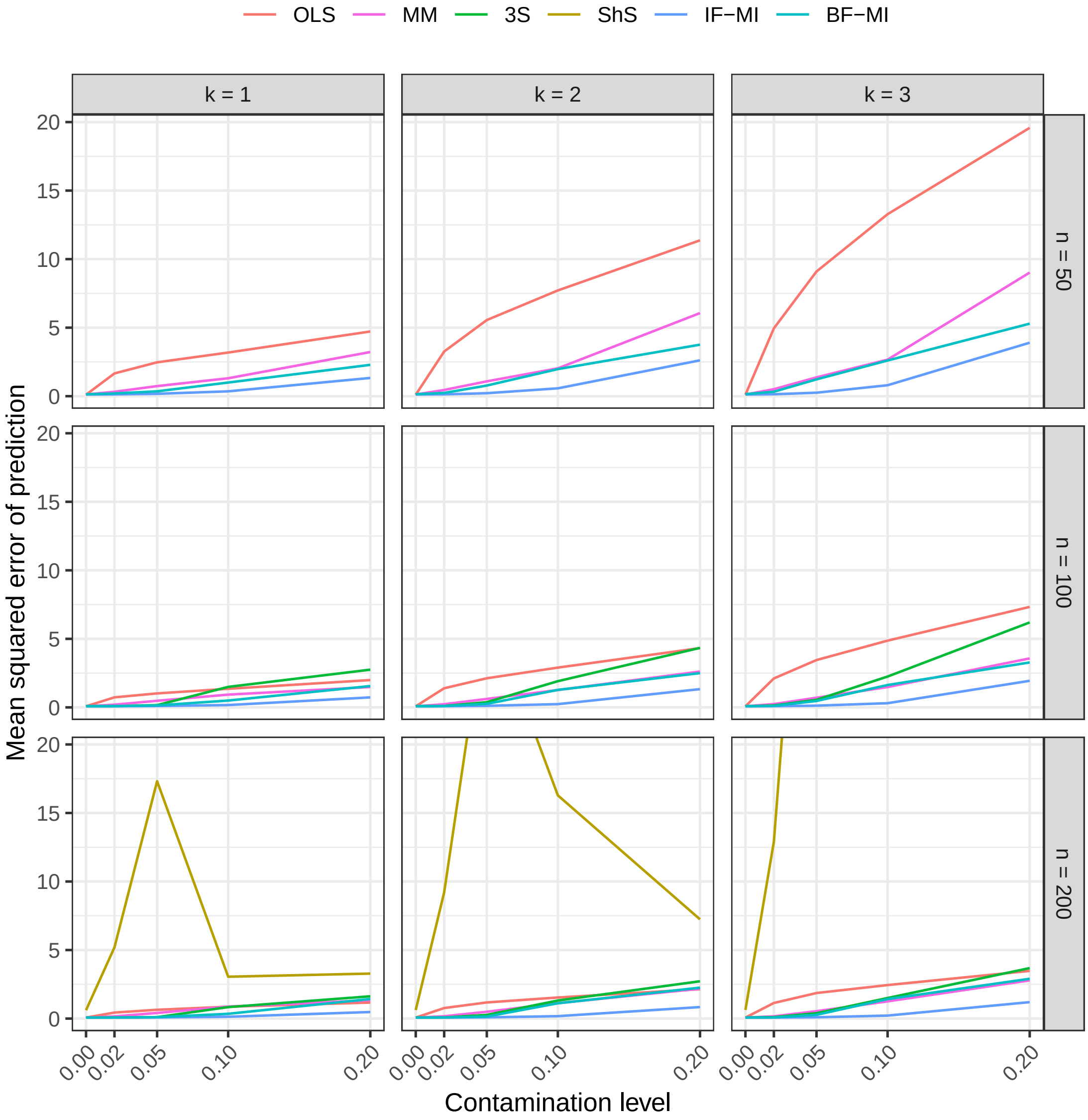

Fig. 10: Results from 1000 simulation runs for the scenario with $D = 20$ compositional parts: the average MSEP for different regression methods is plotted against the contamination level ζ for various sample sizes n and scaling factors k of the covariance matrix in pivot coordinates. Note that the shooting S-estimator (ShS) cannot be applied for $n = 50$ and $n = 100$, as the number of pairwise logratios is larger than the number of observations. In addition, the 3-step regression estimator (3S) is unstable for $n = 50$, yielding an average MSEP that is outside the depicted range on the y-axis.

E Table of results from the sensitivity analysis using the VFA data set

Table 2: Regression coefficient estimates, standard errors and p-values for the VFA data: the proposed compositional MM-regression with a bivariate cellwise outlier filter and multiple imputation with various choices of the parameter τ (which determines the cut-off value in the cellwise outlier filter), with and without variable screening in the imputations.

Without variable screening in imputations

	$\tau = 0.995$			$\tau = 0.99$			$\tau = 0.985$			$\tau = 0.98$			$\tau = 0.975$		
		Std.			Std.			Std.			Std.			Std.	
Variable	Coeff.	Error	p-value	Coeff.	Error	p-value	Coeff.	Error	p-value	Coeff.	Error	p-value	Coeff.	Error	p-value
Intercept	−1.892	2.065	0.361	−2.822	1.998	0.159	−2.367	2.028	0.244	−1.000	2.134	0.637	−1.708	2.115	0.420
$z_1^{(\text{Acetate})}$	0.239	0.105	0.024	0.301	0.084	< 0.001	0.288	0.091	0.002	0.264	0.091	0.004	0.259	0.094	0.006
$z_1^{(\text{Propionate})}$	−0.333	0.061	< 0.001	−0.385	0.054	< 0.001	−0.388	0.056	< 0.001	−0.363	0.056	< 0.001	−0.378	0.057	< 0.001
$z_1^{(\text{Butyrate})}$	0.056	0.061	0.343	0.025	0.050	0.617	0.025	0.053	0.635	0.036	0.054	0.505	0.027	0.054	0.621
$z_1^{(\text{Isobutyrate})}$	0.006	0.057	0.922	−0.014	0.055	0.794	−0.011	0.058	0.847	−0.027	0.067	0.680	−0.039	0.065	0.547
$z_1^{(\text{Isovalerate})}$	0.001	0.036	0.973	0.015	0.034	0.662	0.010	0.037	0.782	0.013	0.041	0.745	0.024	0.040	0.558
$z_1^{(\text{Valerate})}$	0.030	0.058	0.609	0.059	0.064	0.350	0.076	0.064	0.195	0.077	0.075	0.305	0.109	0.076	0.155
ln(ME)	0.738	0.505	0.145	0.755	0.481	0.118	0.664	0.475	0.163	0.448	0.508	0.378	0.622	0.506	0.220
ln(DMI)	−0.358	0.072	< 0.001	−0.397	0.072	< 0.001	−0.370	0.078	< 0.001	−0.352	0.090	< 0.001	−0.373	0.086	< 0.001
ln(Weight)	0.541	0.147	0.004	0.689	0.186	< 0.001	0.653	0.191	< 0.001	0.521	0.198	0.009	0.578	0.194	0.003
$\text{Diet}_{\text{Mixed}}$	0.271	0.046	< 0.001	0.245	0.048	< 0.001	0.241	0.049	< 0.001	0.236	0.053	< 0.001	0.252	0.052	< 0.001

With variable screening in imputations: only variables with absolute robust correlations larger than 0.2 are used

	$\tau = 0.995$			$\tau = 0.99$			$\tau = 0.985$			$\tau = 0.98$			$\tau = 0.975$		
		Std.			Std.			Std.			Std.			Std.	
Variable	Coeff.	Error	p-value	Coeff.	Error	p-value	Coeff.	Error	p-value	Coeff.	Error	p-value	Coeff.	Error	p-value
Intercept	−1.385	2.102	0.510	−2.241	2.029	0.271	−1.661	2.037	0.416	−0.199	2.115	0.925	−0.664	2.145	0.757
$z_1^{(\text{Acetate})}$	0.228	0.105	0.031	0.299	0.084	< 0.001	0.289	0.090	0.002	0.257	0.092	0.005	0.260	0.095	0.007
$z_1^{(\text{Propionate})}$	−0.330	0.061	< 0.001	−0.389	0.055	< 0.001	−0.389	0.056	< 0.001	−0.358	0.057	< 0.001	−0.371	0.057	< 0.001
$z_1^{(\text{Butyrate})}$	0.060	0.060	0.318	0.019	0.050	0.701	0.016	0.053	0.760	0.031	0.056	0.574	0.018	0.056	0.750
$z_1^{(\text{Isobutyrate})}$	0.014	0.057	0.805	−0.004	0.055	0.941	−0.002	0.058	0.974	−0.018	0.065	0.789	−0.033	0.064	0.605
$z_1^{(\text{Isovalerate})}$	0.001	0.036	0.972	0.013	0.034	0.695	0.010	0.037	0.780	0.010	0.041	0.810	0.018	0.040	0.664
$z_1^{(\text{Valerate})}$	0.026	0.058	0.656	0.062	0.063	0.330	0.075	0.065	0.246	0.078	0.075	0.304	0.109	0.076	0.152
ln(ME)	0.667	0.508	0.191	0.702	0.476	0.142	0.585	0.474	0.219	0.348	0.508	0.494	0.497	0.508	0.329
ln(DMI)	−0.341	0.073	< 0.001	−0.367	0.074	< 0.001	−0.345	0.079	< 0.001	−0.320	0.092	< 0.001	−0.334	0.089	< 0.001
ln(Weight)	0.488	0.187	0.010	0.615	0.193	0.002	0.569	0.194	0.004	0.427	0.196	0.030	0.452	0.199	0.025
$\text{Diet}_{\text{Mixed}}$	0.270	0.046	< 0.001	0.243	0.046	< 0.001	0.236	0.048	< 0.001	0.234	0.053	< 0.001	0.250	0.052	< 0.001