Erasmus University Rotterdam
Erasmus School of Economics
Bachelor Thesis Econometrics and Operations Research

Enhancing prediction performance in the context of Cellwise Robust M Regression: a comparison of missing value imputation techniques

Esmée Mulder (523280)

Abstract

Missing values pose significant challenges in the majority of datasets, such as introducing bias or reducing statistical power, necessitating an appropriate imputation approach (Pham, Pandis & White, 2022). Next to this, outliers are encountered often and can lead to similar issues with bias in the model (Osborne & Overbay, 2004). Six missing value imputation methods are analysed to determine the best predictive method in the context of cellwise robust M regression (CRM), which is a robust regression approach that detects and imputes cellwise outliers. In contrast to casewise outliers, which are about an entire observation, cellwise outliers are individual data entries deemed outlying. Looking at cellwise instead of casewise outliers allows for maximal use of non-contaminated data. To address the research question at hand, a comprehensive simulation study is performed, in which the focus is on prediction performance. The main finding is that multiple imputation by chained equations (MICE) with CRM is a potentially good solution. When data are missing completely at random, MICE with CRM performs best in terms of its prediction performance, limiting bias and having the most accurate imputed dataset. Similar conclusions hold when data are missing at random and not missing at random. This research has contributed new insights into the choice of imputation technique in the context of CRM, which can facilitate more accurate and unbiased predictions in the presence of both cellwise outliers and missing values.

Supervisor:	Dr. Aurore Archimbaud
Second assessor:	Dr. Kathrin Gruber
Date final version:	2nd July 2023

The views stated in this thesis are those of the author and not necessarily those of the supervisor, second assessor, Erasmus School of Economics or Erasmus University Rotterdam.

Contents

1 Introduction

Missing values are a part of almost all empirical datasets (Lin & Tsai, 2020; Newgard & Lewis, 2015). Van Buuren (2018) even states that it is "nearly inevitable" that there are no missing values in non-simulated data. Little and Rubin (2019, p. 4) define missing data as the following: "unobserved values that would be meaningful for analysis if observed".

When performing regression analysis, observations with missing values are often simply deleted, which is called listwise deletion (De Souto, Jaskowiak & Costa, 2015). This is because many statistical methods, such as linear regression, only work on complete data (Salgado, Azevedo, Proença & Vieira, 2019). Thus, to perform a regression and thereby estimate the coefficients, either imputing or deleting data is necessary. Yet, the commonly used method of listwise deletion is only suitable when a small percentage of data is missing or if analysing the complete cases does not result in significant bias (Luengo, García & Herrera, 2012). In all other cases, when inappropriately dealing with missing data, for instance by utilising listwise deletion in unsuitable situations, missing values will introduce bias and thus distort conclusions (Donders, Van Der Heijden, Stijnen & Moons, 2006; Groenwold & Dekkers, 2020; Pham et al., 2022; Wulff & Jeppesen, 2017). Another issue stemming from improperly handling missing data in a regression framework is information loss, which becomes especially notable when a large number of observations contain missing values (Zhang, 2016). Also, information loss can consequently lead to reduced statistical power (Pham et al., 2022). Lastly, Madley-Dowd, Hughes, Tilling and Heron (2019) find that missing values reduce efficiency, since listwise deletion decreases sample size which usually leads to incorrect estimates.

There are many potential causes for missing values. Technical errors in the process of data collection are one of these common causes, for instance through incorrect measurements (Lin & Tsai, 2020; Luengo et al., 2012). But, missing values can also arise from non-technical errors like intentional or unintentional survey non-response (Hegde et al., 2019; Newman, 2014).

Outliers are another issue seen time after time in statistical analysis and regression. Many different definitions of outliers exist. Osborne and Overbay (2004, p. 1) for instance define an outlier as "a data point that is far outside the norm for a variable or population". Other definitions tend to revolve more around distance, such as Maddala and Lahiri (1992) who define an outlier as a single observation which is "far removed" from other observations in the data. There are multiple issues as a consequence of outliers. Similar to the problems with missing data, outliers can introduce bias, influence regression estimates and consequently conclusions (Osborne & Overbay, 2004). Therefore, within a classic regression framework, outliers are easily able to affect the least squares estimator and cause unreliable results (Sadouk, Gadi & Essoufi, 2020). Next to this, according to Osborne and Overbay (2004) outliers often intensify error variance and as a result lead to a reduced statistical power.

Nevertheless, not each outlier is the same. A necessary distinction between the different types of outliers (casewise and cellwise) needs to be made. An outlier is casewise if an entire observation is deviating, while cellwise outliers are single outlying cells (one variable of one observation). Looking at outliers on a cell level will help to make optimal use of non-contaminated data, which is preferred to the casewise method, because it can decrease estimation variance. A method that incorporates detecting and imputing cellwise outliers in a robust regression framework has been

developed by Filzmoser, Höppner, Ortner, Serneels and Verdonck (2020), called cellwise robust M regression (CRM). As this method can account for cellwise outliers while simultaneously estimating the coefficients, another method to first handle outliers is no longer necessary. This is important, as "any outlier is only outlying with respect to a model" (Filzmoser et al., 2020, p. 2). However, CRM, similar to other regression techniques, does not adequately handle missing values, since it makes use of listwise deletion.

Therefore, in this paper a variety of missing value imputation (MVI) techniques will be investigated to determine which method is the best in combination with CRM. The following research question will be explored via a simulation study: *What missing value imputation method yields the best prediction performance in combination with cellwise robust M regression?*

Multiple imputation by chained equations (MICE) combined with CRM performs the best when looking at the mean absolute error, mean squared error of prediction and the root mean squared error of imputation, when data are missing completely at random. With different missingness mechanisms (missing at random or not missing at random) the same conclusions can be made for most cases. Besides, as the proportion of missing data increases, almost all six analysed imputation and regression methods perform worse. Lastly, when varying the number of explanatory variables, MICE generally performs better as we have fewer explanatory variables.

The rest of this paper is arranged as follows. In Section 2 the existing literature is discussed, followed by a description of the utilised methods in Section 3. Section 4 provides the results and is followed by an application of the methods on real-life data in Section 5. Lastly, in Section 6, a conclusion and a discussion of limitations and future research suggestions are provided.

2 Literature review

2.1 Types of missing values

Rubin (1976) was the first author to differentiate between the different types of missing data. Currently, the three types are commonly referred to as missing completely at random (MCAR), missing at random (MAR) and not missing at random (NMAR), see Donders et al. (2006).

Data are MCAR if the probability of a single observation of a certain variable being missing, does not depend on the observed data nor on the missing data itself (Lin & Tsai, 2020). Thus, if data are MCAR this indicates that the reason for values being missing is actually random. Therefore, Van Buuren (2018) claims that with MCAR data one can ignore most of the difficulties that otherwise arise from missing values. This makes handling them substantially easier, because most simple imputation techniques should work quite well (Donders et al., 2006).

Graham (2009) defines data that are MAR, as data where the missingness depends on observed data only, so not on the unobserved (or missing) data. Similar definitions are given by Lin and Tsai (2020) and Newman (2014). This implies that in contrast to its name, MAR data are not actually missing at random, because its mechanism is conditional on observed values.

If the missing values depend on the unobserved data (missing values themselves) rather than the observed data, the data are NMAR (Jadhav, Pramod & Ramanathan, 2019; Newman, 2014). Scheffer (2002) finds that for data that is NMAR (or MAR) more sophisticated methods such as multiple imputation are likely necessary.

2.2 Existing literature and hypotheses

A wide variety of MVI methods have been developed and applied in previous research (Lin & Tsai, 2020). This ranges from simple imputation methods like mean, median or mode imputation, to more advanced techniques like multiple imputation (MI) or machine learning approaches, which have been developed in the past couple of decades (Donders et al., 2006). Next to MVI methods, there do exist some regression techniques that can manage missing data. One example of such a technique is the Detect Deviating Cells (DDC) method developed by Rousseeuw and Bossche (2018). DDC utilises the correlations between variables for detecting deviating data cells, such that it can handle cellwise outliers, while simultaneously imputing missing values.

The MVI techniques that will be studied in this research are the following: DDC with ordinary least squares (OLS), DDC with MM estimation (a combination of S and M estimation), stochastic regression imputation with CRM, mean imputation with CRM, an adjusted CRM approach and MICE with CRM. These techniques are described in detail in Section 3.3, while the existing literature on the performance of these methods is discussed here.

To begin, single imputation methods like mean imputation and different variations of regression imputation are simple ad-hoc ways of imputation, and thus come with some advantages. For instance, mean imputation is simple to understand while also being computationally easy and quick (Van Buuren, 2018). Moreover, if the missing data are MCAR, the resulting sample mean should be unbiased (Little & Rubin, 2019). However, according to Donders et al. (2006) and Van Buuren (2018) if data are MAR or NMAR, mean imputation generally produces biased estimates. This can lead to a distorted shape of the distribution if a relatively large percentage of data is missing (Jadhav et al., 2019). Next to this, regression imputation often leads to an overestimation of correlation and an underestimation of variance, even if data are MCAR (Newman, 2014). Nevertheless, simple regression imputation can be improved by adding a random error term to the imputed values, which is called stochastic regression imputation and will be utilised in this paper. Newman (2014) finds that this method can remove the bias generally seen in regression imputation for both MCAR and MAR data. Even though stochastic regression imputation can handle possible bias, all single imputation methods are not equipped to generate accurate standard errors needed for hypothesis testing. Thus, most existing literature concludes that single imputation technique are not optimal (Acock, 2005).

Compared to single imputation, MI is often seen as superior (Kang, 2013; Scheffer, 2002). If data are MCAR or MAR, MI can lead to unbiased estimates as well as accurate standard errors (Donders et al., 2006; Newman, 2014). The reason that MI outperforms single imputation, is that it takes into account uncertainty (Sinharay, Stern & Russell, 2001). Instead of replacing each missing observation with a single value, it reflects this uncertainty as a result of missing values by substituting multiple likely values (Jadhav et al., 2019). The same authors compared different MVI methods on numerical datasets and found that k-nearest neighbour (k-NN) and various MI techniques outperform single imputation all all utilised criteria.

Thus, the expectation is that a profound MI technique like MICE with CRM will perform better than the single imputation methods. Nevertheless, as four of the methods are combined with CRM, and this robust regression technique can handle cellwise outliers, the differences between these methods are expected to be smaller than seen in existing literature. If missing

values are imputed incorrectly, they can still be detected as cellwise outliers and consequently be imputed by CRM, thereby limiting their effect on parameter estimates.

As missing data are almost always either MAR or NMAR and not MCAR, it is expected that the MI method outperforms single imputation techniques (Wulff & Jeppesen, 2017). However, the effects on MCAR data compared to the other missingness mechanisms, are expected to be smaller in this context as it is predicted that CRM can lessen the effects. Furthermore, Madley-Dowd et al. (2019) found that for all different levels of missingness analysed, using MI will benefit the predictions by improving efficiency as well as reducing bias. Moreover, Jadhav et al. (2019) found no large differences between a different percentage of missing values, and they still concluded that k-NN and MI outperform single imputation. Lastly, Kang (2013) finds that even with large datasets MI is a good approach, thus also for the situation when the number of explanatory variables is increased, MICE is expected to outperform single imputation.

The following research will contribute to the previously evaluated literature by analysing if these conclusions also hold when there are also cellwise outliers and making use of CRM. Furthermore, the aim is that with this research more thorough decisions about which imputation method to use when data contains both missing values and cellwise outliers will be possible.

3 Methodology

3.1 Notation

The notation introduced by Filzmoser et al. (2020) will be used here. $\boldsymbol{X} \in \mathbb{R}^{n \times p}$ is the $n \times p$ predictor matrix, with p the number of explanatory variables and n the number of observations. $\boldsymbol{\beta} \in \mathbb{R}^p$ is the $p \times 1$ vector of true regression coefficients and $\boldsymbol{y} \in \mathbb{R}^n$ the $n \times 1$ vector of dependent variables. Moreover, $\boldsymbol{\varepsilon}$ is the $n \times 1$ vector of error terms, which are independent and identically normally distributed. The variables described before, relate to one another via Equation 1.

$$\boldsymbol{y} = \boldsymbol{X}\boldsymbol{\beta} + \boldsymbol{\varepsilon} \tag{1}$$

3.2 Cellwise robust M regression

CRM is an iterative robust regression technique equipped to handle cellwise outliers, for which the algorithm will be described here. Initially, a highly robust estimate against cellwise outliers is applied (such as MM estimation) and is followed by continuously updating these estimates, gaining efficiency. The CRM algorithm as derived from Filzmoser et al. (2020) is iterative and starts with scaling and centering of the data, using the Q_n scale estimator and the L_1 median, which are robust and have a high breakdown point. It is followed by an initialisation step. Using robust MM regression, the initial estimator $\hat{\boldsymbol{\beta}}$ is obtained from the original $\mathbf{x}_i$ and y_i. Following this, utilising this initial estimator $\hat{\boldsymbol{\beta}}$, Algorithm 1 is performed. As can be seen in step 8, $\mathbf{X}_\omega$ and $\mathbf{y}_\omega$ are obtained and this is followed in step 9 by using this imputed and weighted data to calculate a new estimator $\hat{\boldsymbol{\beta}}$ with least squares. Algorithm 1 is then repeated with this updated $\hat{\boldsymbol{\beta}}$ estimator and the updated data $\mathbf{X}_\omega$ and $\mathbf{y}_\omega$, while saving the previous estimator $\hat{\boldsymbol{\beta}}$. Algorithm 1 is repeated until $\text{mean}(|\hat{\boldsymbol{\beta}} - \hat{\boldsymbol{\beta}}_{\text{previous}}|) < 0.01$, where $\hat{\boldsymbol{\beta}}_{\text{previous}}$ is the estimator from the previous iteration.

Algorithm 1 Cellwise robust M regression

1: Calculate the residuals based on the estimator $\hat{\boldsymbol{\beta}}$ via the following: $r_i = y_i - \mathbf{x}_i^T\hat{\boldsymbol{\beta}}$ for all $i \in \{1, ..., n\}$.
2: An observation is detected as a casewise outlier if: $\frac{|r_i|}{1.4826 \times \text{median}_i|r_i|} > z_{0.95}$, where $z_{0.95}$ is the 0.95 quantile of the standard normal distribution.
3: For each case that is deemed outlying apply the SPADIMO algorithm to determine which are cellwise outliers.
4: If not all variables in the casewise outlier contribute to the outlyingness, impute each outlying cell using Algorithm 2.
5: After applying Algorithm 2 to each outlying cell, denote the new data matrix by $\tilde{\mathbf{X}}$.
6: Use $\tilde{\mathbf{X}}$ to update the residuals (see step 1): $\tilde{r}_i = y_i - \tilde{\mathbf{x}}_i^T\hat{\boldsymbol{\beta}}$ for all $i \in \{1, ..., n\}$.
7: Use the Hampel weight function to calculate the case weights: $\omega_i = w_H(\frac{|\tilde{r}_i|}{1.4826 \times \text{median}_i|\tilde{r}_i|})$.
8: Let $\Omega = \text{Diag}(\sqrt{\omega_1}, ..., \sqrt{\omega_n})$ be the diagonal matrix with the square root of the case weights as diagonal elements and update the imputed data as follows: $\mathbf{X}_\omega = \Omega\tilde{\mathbf{X}}$ and $\mathbf{y}_\omega = \Omega\mathbf{y}$.
9: Obtain the estimator $\hat{\boldsymbol{\beta}}$ from a least squares regression on the imputed and weighted data $\mathbf{X}_\omega$ and $\mathbf{y}_\omega$ and save $\hat{\boldsymbol{\beta}}$.

Looking at step 2 of Algorithm 1, one can observe that the residual calculated from the estimator $\hat{\boldsymbol{\beta}}$ is utilised to detect outliers on a case level. If the absolute standardized residual is greater than the 95% quantile of the standard normal distribution (1.645), an observation is deemed a casewise outlier. The SPADIMO (SPArse DIrections of Maximal Outlyingness) algorithm developed by Debruyne, Höppner, Serneels and Verdonck (2019) is then used to differentiate within one observation which cells contribute to the outlyingness. If a cell is flagged by SPADIMO to be outlying, it is imputed via its two nearest neighbours in Algorithm 2. Taking only the clean cells, the mean of the two nearest neighbours is calculated for each outlying cell (step 4 of Algorithm 2). The newly imputed dataset $\tilde{\mathbf{X}}$ from step 5 is then used to update the residuals in step 6. Following this the Hampel redescending weight function, is used to calculate case weights, which downweight the casewise outliers in step 8. This Hampel function is chosen because of its good trade-off between robustness and efficiency, like with MM estimation (Filzmoser et al., 2020). The algorithm stops if the mean absolute difference between the previous and new regression estimates is small enough, in this case smaller than 0.01.

Algorithm 2 Outlier imputation in CRM

1: For each outlying case $\mathbf{x}_i$ with index i, set q as the number of variables detected as cellwise outliers by SPADIMO in $\mathbf{x}_i$ and p as the total variables in $\mathbf{x}_i$.
2: Let C be the set of $q < p$ outlying variables in $\mathbf{x}_i$.
3: Detect the two nearest neighbours $\mathbf{x}_{n_1}$ and $\mathbf{x}_{n_2}$ of outlier $\mathbf{x}_i$ in the subspace $\{1, ...p\} \setminus C$, only among observations that are not outlying.
4: Impute each outlying cell by the mean of its two nearest neighbours: $x_{iq} = (x_{n_1q} + x_{n_2q})/2$.

This outlier imputation algorithm is a version of other existing nearest neighbour (NN) imputation methods. NN techniques, such as k-NN are efficient in filling out missing values using related cases that are in the rest of the dataset (Beretta & Santaniello, 2016).

Several evaluation methods are used to analyse model performance, following the example of Filzmoser et al. (2020). The methods are: mean absolute error (MAE) to measure bias, mean squared error of prediction (MSEP) to evaluate prediction performance, root mean squared error

of imputation (RMSEI) to measure the accuracy of imputation, and the precision and recall. Define n_c as the number of clean cells, with $n_c < n$ and I as the set of uncontaminated cases. Also, $\hat{\beta}_j$ and $\hat{y}_i$ are the predicted values, while β_j and y_i are the actual values. The formula for the MAE is given in Equation 2 and that of the MSEP is given in Equation 3. $\mathbf{X}^{imp}$ is the imputed matrix, as imputed by either DDC or one of the other MVI methods. Using $\mathbf{X}^{imp}$ and the original $\mathbf{X}$ the RMSEI can be calculated using Equation 4.

$$\text{MAE} = \frac{1}{p}\sum_{j=1}^{p}|\hat{\beta}_j - \beta_j| \tag{2}$$

$$\text{MSEP} = \frac{1}{n_c}\sum_{i\in I}(\hat{y}_i - y_i)^2 \tag{3}$$

$$\text{RMSEI}(\mathbf{X}^{imp}, \mathbf{X}) = \sqrt{\frac{1}{np}\sum_{i=1}^{n}\sum_{j=1}^{p}(x_{ij}^{imp} - x_{ij})^2} \tag{4}$$

For the precision and recall (Equation 5), the following are defined: a_{12} is the number of cells that are actually outliers, but not detected by CRM, while a_{21} is the number of cells that are not outliers, but are flagged as such. a_{22} is the number of correctly identified outliers. The precision gives the ratio of correctly flagged cellwise outliers over the total number of cells flagged as outliers and the recall is the ratio of correctly flagged cellwise outliers over all actual outliers.

$$\text{Precision} = \frac{a_{22}}{a_{21} + a_{22}} \;;\; \text{Recall} = \frac{a_{22}}{a_{12} + a_{22}} \tag{5}$$

3.3 Missing value imputation techniques

3.3.1 Mean imputation

Mean imputation is one of the most commonly applied methods to handle missing data (Jadhav et al., 2019; Lin & Tsai, 2020). It simply replaces the missing value with the sample mean of the variable, see Equation 6. y_{ij} is the missing value of variable y_j for observation i and $\bar{y}_j^{-(j)}$ is the mean of y_j excluding the missing values (Little & Rubin, 2019). After all missing values are imputed, CRM is performed on the imputed dataset as usual.

$$y_{ij} = \bar{y}_j^{-(j)} \tag{6}$$

3.3.2 Detect Deviating Cells (DDC)

Rousseeuw and Bossche (2018) proposed the Detect Deviating Cells (DDC) method which can detect and predict the values of casewise and cellwise outliers, while simultaneously dealing with missing values. Both OLS and MM estimation are combined with DDC, to predict the model. The data are first standardized with a robust estimator for location as well as scale. Next to this, the cells that stand out in reference to their column, are flagged. Each missing cell as well as each detected outlier is then predicted, based on only the unflagged cells of the row of the column that is correlated with the column which contained the flagged cell (Rousseeuw &

Bossche, 2018). After DDC is performed, MM or OLS regression is applied on the imputed data matrix to predict the regression estimates.

3.3.3 Adjusted CRM

This method is constructed such that CRM handles missing values via its own way of handling cellwise outliers. To impute missing values via the CRM algorithm, the cells will be temporarily 'imputed' using mean imputation, but then artificially made an outlier. Mean imputation is described earlier in Section 3.3.1. To make these imputed values outlying, the mean will be multiplied by 10 (Equation 7). Following this, CRM is applied, and it is expected that these fabricated outliers are detected by SPADIMO and imputed by the NN method in Algorithm 2.

$$y_{ij} = 10 \times \bar{y}_j^{-(j)} \tag{7}$$

3.3.4 Stochastic regression imputation

Lin and Tsai (2020) find that linear regression as an imputation method is also commonly used. Traditional linear regression imputation is generally performed in two steps. First, only making use of the observed (non-missing) data a regression model is estimated. Afterwards, this model is used to predict and replace the missing values. Stochastic regression imputation is a more refined approach, which can address problems with bias that arise in linear regression, by adding noise to the prediction (Heymans & Eekhout, 2019). This method adds a random draw from the normal distribution to the predicted values from the linear model (Van Buuren, 2018). Denote X_{mis} as the subset of observations of $\mathbf{X}$ that contains a missing value in variable y. y_{imp} denotes the vector of imputed values, which is a subset of y. The stochastic regression imputation follows Equation 8, where ε_{imp} is randomly drawn from normal distribution as $\varepsilon_{imp} \sim N(0, \hat{\sigma}^2)$, and $\hat{\beta}_0$ and $\hat{\beta}_1$ are the least squares estimates calculated over the observed data. This will be performed in R using the *mice* function from the package *mice*, developed by Van Buuren and Groothuis-Oudshoorn (2011). However, one performs single imputation with this function by setting the number of generated datasets and the maximum number of iterations to 1, while setting the method to *norm.nob* which invokes stochastic regression imputation.

$$y_{imp} = \hat{\beta}_0 + X_{mis}\hat{\beta}_1 + \varepsilon_{imp} \tag{8}$$

3.3.5 Multiple imputation by chained equations (MICE)

MI is becoming an increasingly popular technique for dealing with missing values (Jadhav et al., 2019). Single imputation methods replace a single value for each missing observation, thereby not taking into account possible uncertainty. MI on the other hand, generates m different datasets which are copies of the original data, and in each one missing values are imputed. MI is therefore able to provide unbiased estimates and accurate standard errors for data that is both MCAR and MAR (Newman, 2014). A chain of regression equations is utilised to get the imputations, meaning missing values are imputed one by one instead of all simultaneously, using all available information on other variables (Heymans & Eekhout, 2019). The algorithm used for MICE in

this paper is presented in Algorithm 3. First, m copies of the existing dataset are made. For each copy an iterative predictive mean matching (pmm) procedure is performed, which is given in Algorithm 4 until each missing value is imputed. This procedure for each dataset copy is repeated T times (the maximum number of iterations), updating the imputed values each time. Finally, the m imputed versions are pooled together by taking the mean. CRM is applied to this final dataset, to impute outliers and produce a model fit.

Algorithm 3 Multiple imputation by chained equations

1: Create m duplicates of the incomplete dataset $\mathbf{X}$.
2: From left to right, for each variable in $\mathbf{X}$ containing missing values perform Algorithm 4.
3: Perform T iterations of step 2 for each of the m datasets.
4: Extract the m imputed datasets and take the mean of the different versions.
5: Apply CRM on the final (mean) imputed dataset.

Define y as the target variable, which is a single variable containing missing values. For each y, n_m is the number of observations containing a missing value and n_o the number of observed observations, such that $n = n_m + n_o$. Then y_o is a $n_o \times 1$ vector, which is the subset of observed y observations. Define X as the matrix of remaining explanatory variables and divide into: X_o (a $n_o \times p$ matrix containing the rows of y_o) and X_m (a $n_m \times p$ matrix containing the rows of y_m). The Algorithm for pmm (Algorithm 4), is derived from Rubin (2004) and Van Buuren (2018).

Algorithm 4 Predictive mean matching

1: Define the matrix of cross-products as follows: $C = X_o^{'} X_o$.
2: Take some small paramter κ (1e-0.5) to calculate $K = (C + diag(C)\kappa)^{-}1$, where $diag(C)$ contains the diagonal elements of C.
3: Calculate the regression weights as follows: $\hat{\beta} = KX_o^{'} y_o$.
4: Calculate the degrees of freedom, as the number of variables that are observed minus the number of predictor variables $f = n_o - p$, and use this to draw a random variable from the chi-squared distribution $\dot{r} \sim \chi^2_f$.
5: Calculate the standard errors as follows: $\dot{\sigma} = (y_o - X_o\hat{\beta})^{'}(y_o - X_o\hat{\beta})/\dot{r}$.
6: Use a Cholesky decomposition to obtain $K^{1/2}$.
7: Draw p independent $N(0,1)$ variables and save them in the vector $\dot{z}$.
8: Calculate $\dot{\beta} = \hat{\beta} + \dot{\sigma}\dot{z}K^{1/2}$ using all previously obtained results.
9: Calculate $\dot{\eta}(i,j) = |X_{o,i}\hat{\beta} - X_{m,j}\dot{\beta}|$ where $i = 1, ..., n_o$ and $j = 1, ..., n_m$.
10: Construct n_m sets S_j that each contain d candidate donors, such that $\min_j \sum_d \dot{\eta}(i,j)$ for $j = 1, ..., n_m$.
11: For each $j = 1, ..., n_m$ draw one d_j randomly from S_j and get the imputed value $\dot{y}_j = y_{d_j}$.

In step 8 the value that is randomly drawn from the posterior distribution of $\hat{\beta}$ is calculated, using the results from steps 1 to 7.The distance between the predicted value of y_o and the drawn value of y_m is calculated in step 9. Then for each missing entry the d donor candidates for which this distance is minimized are found. In step 11 one of these donors is randomly chosen and used to impute the missing value. To implement MICE with pmm the function *mice* in R is used, but now for a MI approach as opposed to single imputation in Section 3.3.4. Wulff and Jeppesen (2017) find that pmm imputation accurately resembles the actual values.

In the pmm algorithm applied in MICE, the visit order is set from left to right, which should not affect the imputation as long as each variable is visited often enough (Van Buuren,

Brand, Groothuis-Oudshoorn & Rubin, 2006). The number of visitations of Algorithm 4 for each dataset is determined by the maximum number of iterations T. Wulff and Jeppesen (2017) suggest starting at $T = 10$, while other existing research more commonly uses $T = 20$, for instance White, Royston and Wood (2011). However, as more iterations also reduce efficiency, this results in a trade-off and therefore $T = 10$ is chosen. Furthermore, the number of imputed datasets, m needs to be chosen. Although usually $m = 5$ is suggested, White et al. (2011) have proposed a new rule of thumb which can limit the loss of power, namely that m should be at least as large as the percentage of incomplete observations. This rule of thumb is based on the original research by Graham, Olchowski and Gilreath (2007) who first challenged the norm of $m = 5$. Other suggestions, for instance by Azur, Stuart, Frangakis and Leaf (2011) make the case that $m = 40$ leads to the most accurate results, but this is often impractical because of its effect on running time. Thus, the rule of thumb by White et al. (2011) is used for a good trade-off between accuracy and efficiency.

4 Results

4.1 Performance of cellwise robust M regression without missing values

The presence of cellwise outliers necessitates a robust regression approach. In order to find out if CRM outperforms other (robust) regression techniques it is compared to MM, DDC-MM, OLS and DDC-OLS. To compare the methods a simulation study is performed on data of $n = 400$ observations and $p = 50$ explanatory variables constructed from draws of the p-dimensional multivariate normal distribution. Afterwards, 5% casewise and 10% cellwise outliers are added, where a cellwise outlier is constructed by adding $k = 6$ times the standard deviation of the variable to its mean, plus a random draw from the standard normal distribution. The simulation is repeated 100 times and the results are averaged. The values displayed at the bottom of the boxplots, give the mean value for each method (also given by the red dots in each boxplot), with the optimal values displayed in bold.

CRM performs best in terms of limiting bias and having the best prediction performance compared to the four other methods. Furthermore, compared to DDC, CRM has a more accurate imputed data matrix. These results and other results obtained are exactly the same as in the original research by Filzmoser et al. (2020) and are displayed in their entirety in Appendix A.

In the baseline simulation study with $p = 50$ variables, CRM was seen to outperform the other (robust) regression methods on all evaluation criteria. However, it is also interesting to see if this conclusion also holds for other levels of p. Thus, the same simulation procedure is performed, but extended such that there are a different number of explanatory variables. Each of the five regression techniques is analysed across $p \in (10, 25, 50, 100, 200, 300)$ and the simulation is repeated 50 times for each p.

Figure 1, shows that as p increases, the average MAE for all five regression methods seems to converge. As p is small (10 or 25) the differences are large between the various methods, while as p is large (300), especially CRM, MM, and OLS perform on a comparable level, while DDC-OLS and DDC-MM deviate slightly (downwards or upwards).

The average MSEP (left) and RMSEI (right) are depicted in Figure 2. Looking at the MSEP,

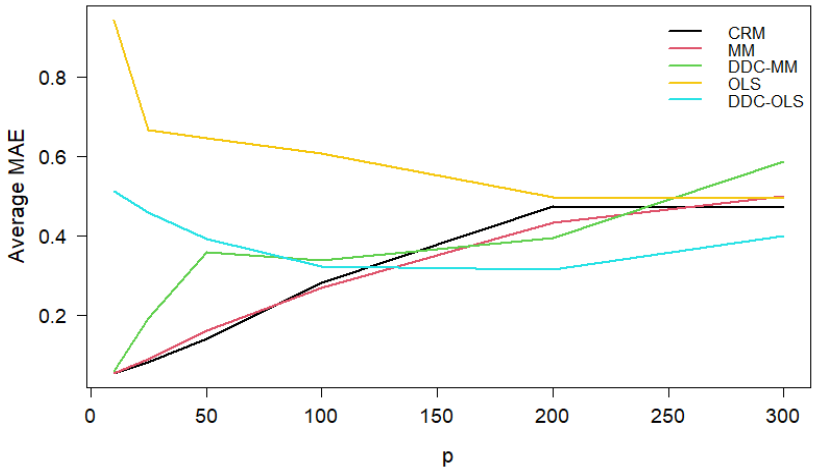

Figure 1: Average MAE for the five regression methods for different levels of p.

as p is large, CRM, MM and OLS perform comparably, while the MSEP of DDC-OLS is slightly higher and that of DDC-MM is substantially higher. Moreover, the average MSEP of CRM, MM and DDC-MM remains relatively stable as p varies, while for OLS and DDC-OLS the average MSEP decreases as p increases. This indicates that the prediction performance of the OLS and DDC-OLS methods improves with more explanatory variables, while the other methods are not really affected by the level of p. For the average RMSEI, a clear difference between CRM and DDC can be observed. The quality of the imputed data matrix of DDC is stable, while CRM is highly influenced by p. The imputed matrix of CRM is more accurate when p is smaller, but DDC is better with larger p.

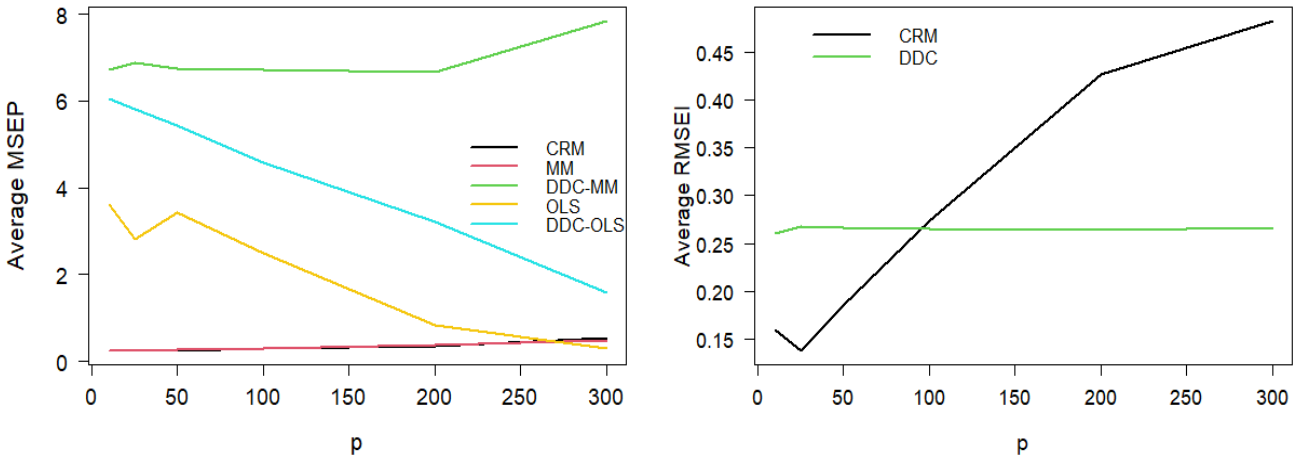

Figure 2: Average MSEP (left) and RMSEI for CRM and DDC (right), for different levels of p.

Lastly, the average precision and recall of CRM for a varying p are given in Figure 3. Both the precision and recall generally decrease as p increases. This indicates that with more explanatory variables, CRM becomes increasingly worse at correctly identifying outlying cells. CRM selects too many cells as outliers and consequently imputes them, increasing bias, which could also be seen in Figure 1 by the increase in the MAE. Moreover, with more explanatory variables CRM also misses substantially more outlying cells and therefore does not impute them, leading to a less accurate imputed data matrix, which is confirmed by the RMSEI in Figure 2.

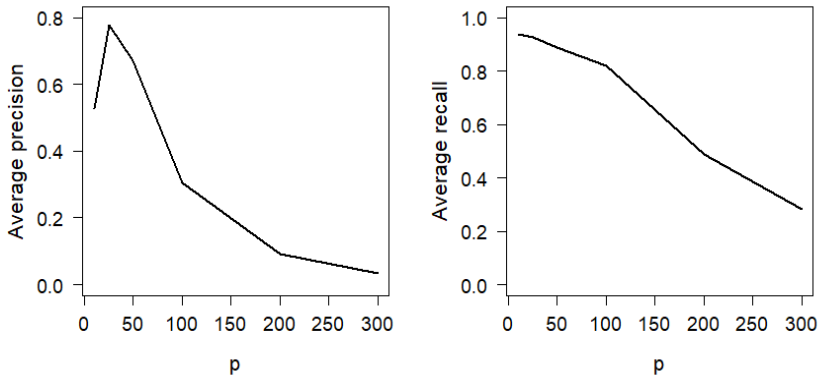

Figure 3: Average precision (left) and average recall (right) of CRM for different values of p.

4.2 Missing value imputation on MCAR data

The MVI methods as discussed in Section 3.3 are applied here. The simulation is constructed just as before (with $n = 400, p = 50$, $k = 6$, 5% casewise outliers and 10% cellwise outliers), however, before contamination is added missing values are created. In this baseline analysis, 10% missing values are used and the data are MCAR. This missingness mechanism is applied using the *delete_MCAR* function from the package *missMethods* in R. Missing values will be allowed to occur in each column. Also, the rule of thumb for MICE imputation will be applied here, such that the number of imputed datasets m is set to 10, and T is also set to 10.

Figure 4 shows the MAE, MSEP and RMSEI. From this it can be seen that MICE with CRM performs best, as it has the lowest average MAE, MSEP and RMSEI. Furthermore, the robust regression method which is designed to deal with missing values as well as cellwise outliers, DDC, does not seem to outperform MICE, but does have a smaller interquartile range for the criterion that measures the accuracy of imputation (RMSEI).

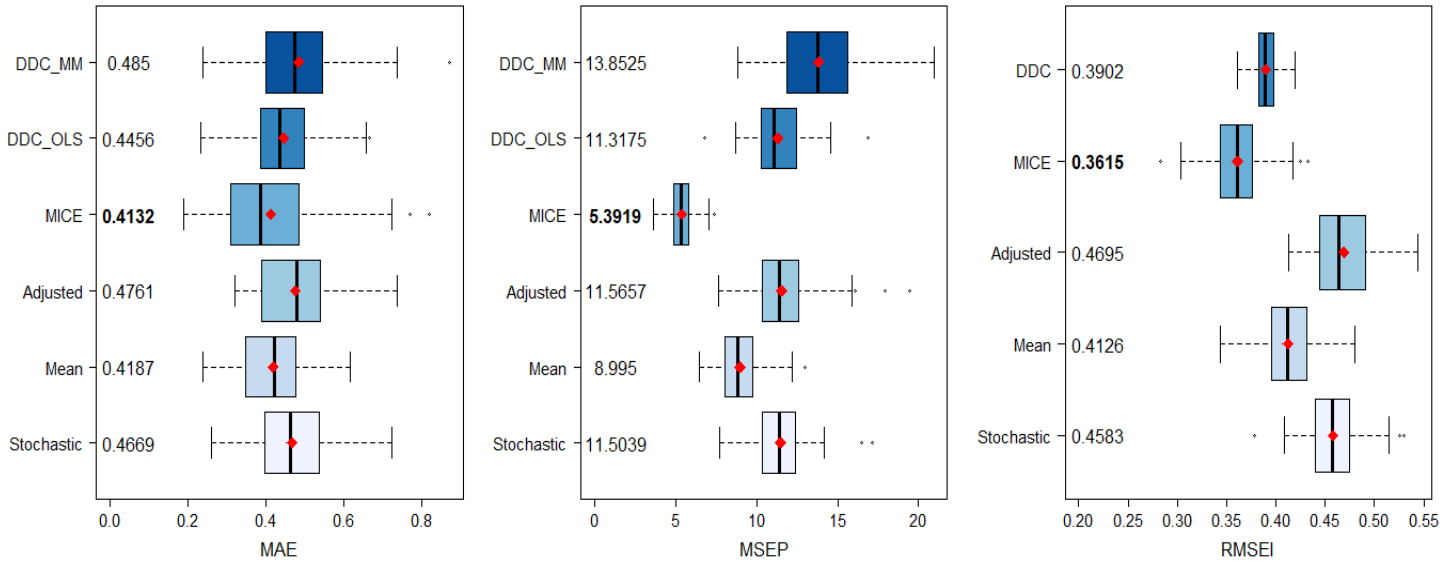

Figure 4: Boxplot of MAE (left), MSEP (centre) and RMSEI (right) across different MVI methods.

It is also interesting to see how mean imputation compares to the other MVI methods in Figure 4. In the existing literature, mean imputation is seen as a poor-performing MVI technique like by Acock (2005), but if mean imputation is combined with the robust estimator from CRM

it performs similarly to for example stochastic regression imputation. Finally, the adjusted CRM method, which utilizes the NN technique for outlier imputation in the CRM algorithm to deal with missing values, performs worse than the mean imputation CRM method for all criteria.

In Figure 5 the boxplots of precision and recall of the four MVI methods that use CRM are depicted. It can be seen that the precision and recall of mean imputation and adjusted CRM are similar. Both show a high precision but average recall, implying that if these methods select a cell to be an outlier, it almost always is, but both methods also miss quite some outliers. Furthermore, stochastic regression imputation results in almost equal precision and recall. Lastly, MICE with CRM has on average low precision, but the highest recall of all four methods. This indicates that although MICE with CRM detects a substantial number of the actual outliers, it also flags too many cells as outliers, thereby incorrectly imputing too many values. Surprisingly, the low precision of MICE resulting in many unnecessary imputations did not substantially affect the bias as indicating by the lowest average MAE in Figure 4.

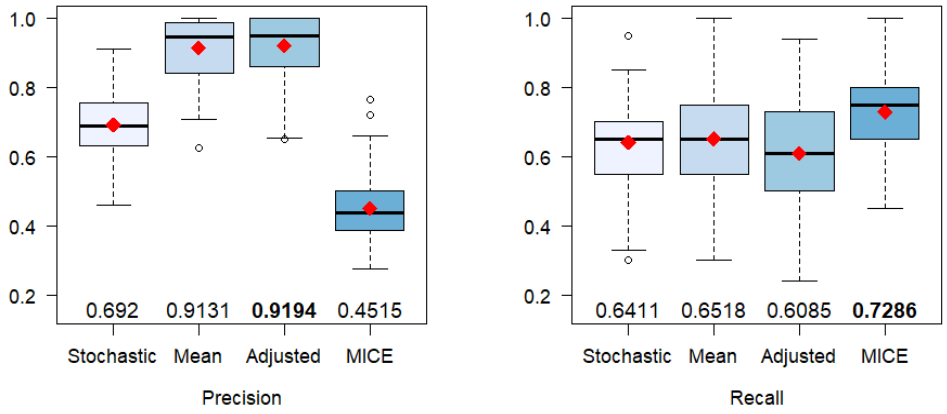

Figure 5: Boxplot of precision (left) and recall (right) of the four imputation methods with CRM.

All in all, when looking at the precision and recall none of the four methods seems to be performing substantially better than the others. Stochastic regression imputation with CRM has the best trade-off between precision and recall, while mean imputation with CRM and adjusted CRM have on average high precision with a relatively low recall. Lastly, MICE has the lowest precision, but the highest recall. Thus, even though MICE performed best on the other three evaluation criteria, if the precision is deemed crucial, other MVI techniques (like mean imputation) might be more suitable.

4.3 Differences and similarities between MCAR, MAR and NMAR data

The assumption that data are MCAR is almost always unrealistic for real-life data (Newman, 2014; Van Buuren, 2018). Most missing data are actually MAR or NMAR (Donders et al., 2006). Therefore, the same simulation as in Section 4.2 is performed, but now the missingness mechanism is MAR or NMAR. Missing values are created in R using the functions *delete_MAR_rank* and *delete_MNAR_rank*. The full results of the MAE, MSEP, RMSEI and precision and recall for the MAR and NMAR cases can be found in Appendix B and C. Here the differences and similarities will be shortly discussed. Overall, for all missingness mechanisms, similar results can be seen. MICE generally outperforms all other methods, like in the MCAR case, except for the

MAE with MAR data. Also, the precision and recall follow similar patterns for the three missingness mechanisms. However, for MAR data, DDC-OLS has the lowest average MAE and not MICE, as with MCAR and NMAR. A comparison of the average MAE of MICE and DDC-OLS across the three missingness mechanisms can be seen in Figure 6. Generally, the performance of MICE and DDC-OLS does not differ substantially. It is thus unsurprising that DDC-OLS might sometimes perform better than MICE, leading to less biased estimates.

Another difference between the missingness mechanisms is surrounding adjusted CRM. It can be seen in Figure 7, that adjusted CRM performs quite similarly for MCAR and MAR data, but when data are NMAR adjusted CRM has a clearly higher MAE, MSEP and RMSEI. Thus, adjusted CRM does not seem to be a good method for NMAR data.

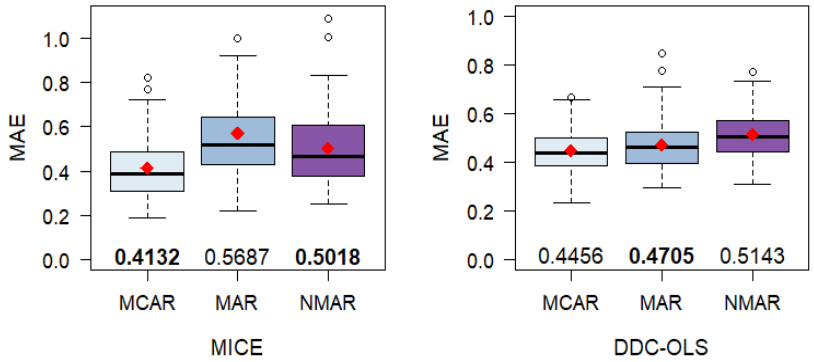

Figure 6: Boxplot of the MAE of MICE (left) and DDC-OLS (right) for different missingness mechanisms.

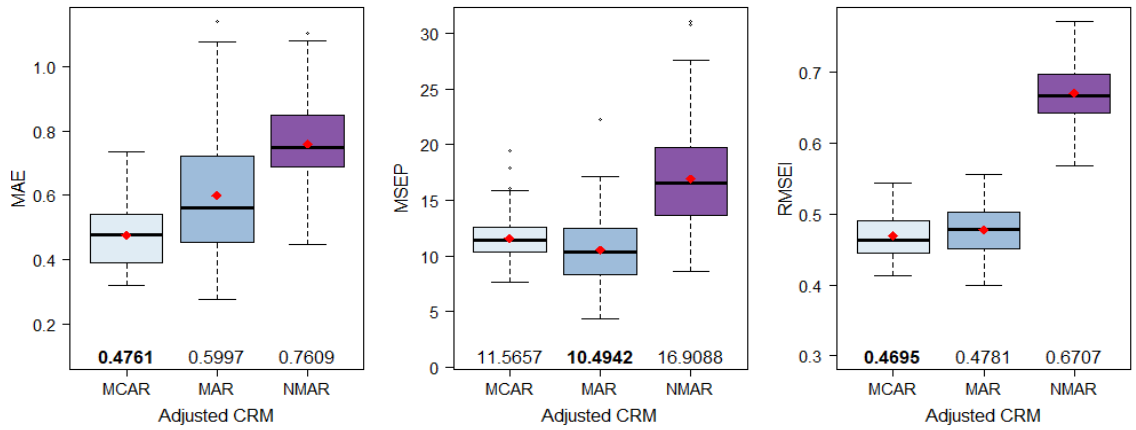

Figure 7: Boxplot of MAE (left), MSEP (centre) and RMSEI (right) of Adjusted CRM for different missingness mechanisms.

4.4 Different percentage of missing values

In the following section, the simulation as in Section 4.2 is again performed. Yet, in this case the percentage of missing values (r in this case) is varied across $r = (10, 20, 30, 40, 50)$. Even though Madley-Dowd et al. (2019) find that the proportion of missing values should not be used to determine which imputation method to use, it is still important to see the performance of the MVI methods across a wide range of missing rates. For each value of r the simulation is repeated 50 times and the results are again averaged.

In Figure 8 the average MAE is depicted. The first interesting observation is that when r increases, all six imputation and regression methods have a higher average MAE, indicating more biased estimated coefficients. Furthermore, stochastic regression imputation with CRM, mean imputation with CRM and adjusted CRM perform similarly across the different levels of r. MICE imputation on the other hand, shows a large increase in the average MAE when r goes from 10% to 30%. But then as r increases to 40% this does not persist. Lastly, the average MAE for DDC-MM and DDC-OLS increase similarly to most other methods, but as r becomes much larger (30% to 50%) there is a large increase in the average MAE.

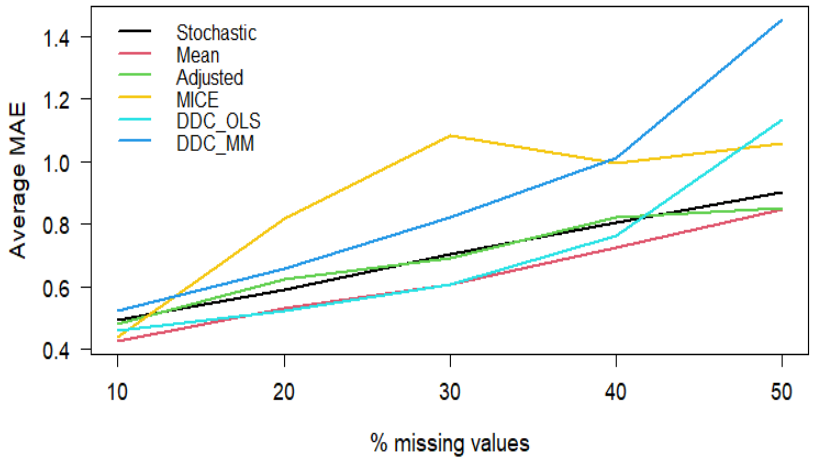

Figure 8: Average MAE for the different MVI methods for a different percentage of missing values.

The average MSEP and RMSEI depict a similar overall trend, with increasing average values with a higher percentage of missing values, as generally seen in with the MAE. Both of these evaluation criteria have a linear trend and they can be seen in Appendix D.

In Figure 9, the average precision (left) and recall (right) for the four imputation techniques that are combined with CRM are depicted. Mean imputation and adjusted CRM perform similarly for the average precision. For both methods the average precision stays roughly the same until around 40% missing values but then drastically decreases until 50%. On the other hand, both stochastic regression imputation and MICE show a much more steady trend. The average precision of stochastic regression imputation with CRM only decreases slightly as r increases, while the average precision of MICE with CRM first decreases until 30% missing values and then somewhat increases to 50%. The right panel of Figure 9 shows the average recall, telling a different story. For all four methods, the average recall steadily decreases with an increase in the percentage of missing values. No substantial differences between the methods can be seen, except that MICE has the highest average recall for all levels of missing values. The decreasing precision and recall for increasing r is to be expected when looking at Figure 8, where the increasing MAE shows that the bias also increases for larger r.

In conclusion, for all methods analysed it can be seen that for both the MAE and average precision and recall the MVI and robust regression techniques perform better with less missing values. They are more effective at limiting bias and correctly identifying outliers, and all methods become increasingly worse at these aspects with larger percentages of missing values.

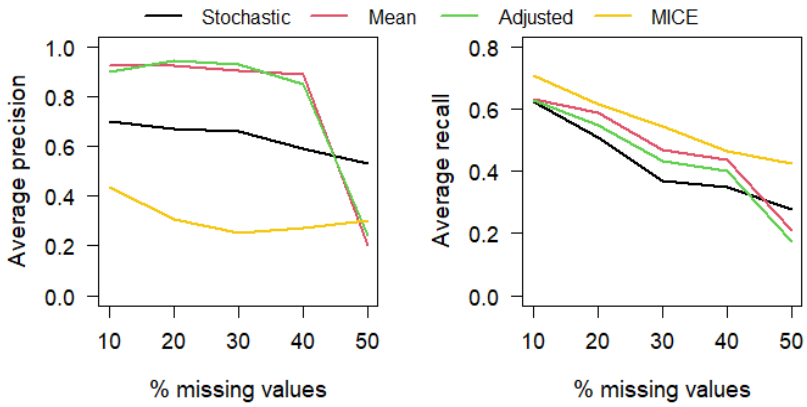

Figure 9: Average precision (left) and average recall (right) of the four imputation methods with CRM.

4.5 Comparing the performance of two MICE variations

The same analysis as in Section 4.4 is performed, also with a variation in the MICE algorithm. Instead of varying the number of datasets (m) for MICE according to the percentage of missing values, m is now set to 5, which is another common suggestion in existing research (White et al., 2011). The differences between the performance of the two MICE variations on the MAE, MSEP and RMSEI criteria are in Appendix E. Figure 10 shows the average precision and recall for the two variations of MICE with CRM. One can see that the general shape of both the precision and recall are comparable. The average precision is again a parabolic curve. The average recall for both situations is comparably decreasing. All in all, no large differences between the two variations can be found, in contrast to existing literature (Azur et al., 2011). This contrasting finding is important, as a larger m is also more computationally expensive, and without the expected benefits it should therefore not always be recommended in the context of CRM.

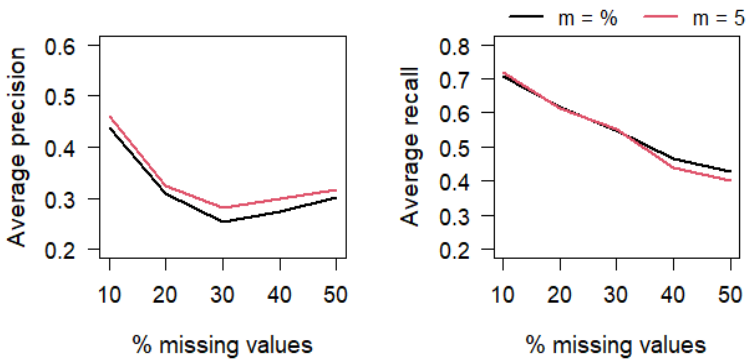

Figure 10: Average precision (left) and average recall (right) of MICE with CRM for two variations.

4.6 Different number of explanatory variables

Here the same simulation study as in Section 4.2 is performed, but now the number of explanatory variables (p) is varied. White et al. (2011) claim that imputation performance might be harmed if too many variables are added to the model, while Wulff and Jeppesen (2017) find that for instance applying MICE on small datasets can also be an issue. Besides, Azur et al. (2011) claim that too many explanatory variables might be impractical, because of the effects on running time. The set analysed is $p \in (10, 25, 50, 75, 100)$ and for each p the simulation is

repeated 50 times and the results are averaged.

To begin, the average MAE for each of the six MVI techniques is given in Figure 11. The most notable observation is that MICE with CRM is evidently different from the other five methods. As p increases, the average MAE also increases quite drastically, especially from $p = 50$ to $p = 100$ Therefore, as the number of explanatory variables increases, the other methods result in less biased coefficients. On the contrary, if p is small MICE performs much better. The remaining five methods follow a comparable pattern, with no large differences as p increases, indicating similar effects on the bias of regression coefficients.

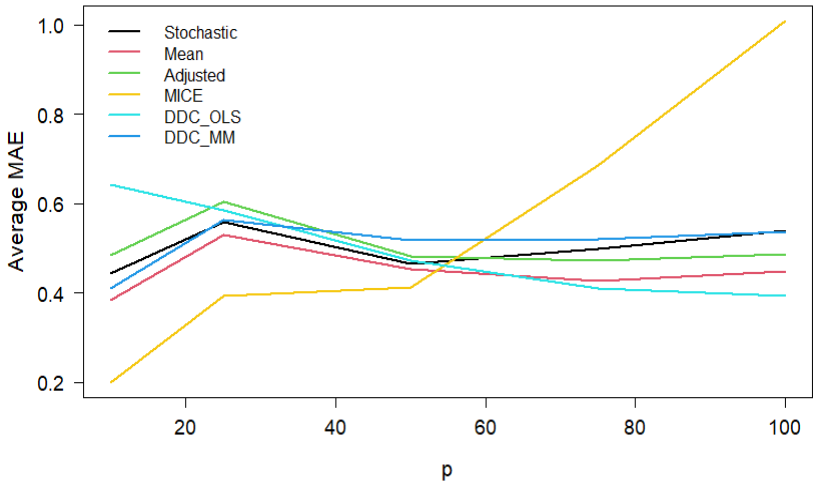

Figure 11: Average MAE for the different MVI methods for various levels of p.

The average MSEP as seen in the left panel of Figure 12 however, tells a completely different story. MICE with CRM has the lowest average MSEP irrespective of the number of explanatory variables. Furthermore, DDC-MM, adjusted CRM, stochastic regression imputation with CRM and mean imputation with CRM all show a similar pattern. These four methods show a slight decrease in the average MSEP going from $p = 10$ until around $p = 50$, but as p increases further they all have an increased average MSEP. The only method that does not share this pattern is DDC-OLS, which has a decreasing average MSEP with increasing p. This indicates that the prediction performance of DDC-OLS decreases with more explanatory variables. To continue, the average RMSEI is given in the right panel of Figure 12. All imputation techniques show a similar trend, except for DDC, which has an almost constant average RMSEI across p. The remaining four methods all show a slight decrease in their average RMSEI from $p = 10$ to $p = 25$, however as p increases further all have an increased RMSEI again. Thus, it can be seen that the quality of the imputed matrix is highest for MICE when there are only a small number of explanatory variables, but at some point, DDC becomes more accurate in imputing the data.

Lastly, in Figure 13 the average precision and recall for the four imputation methods combined with CRM are given for the different levels of p. Similarly to the situation with a varying percentage of missing values in Figure 9, both mean imputation and adjusted CRM follow roughly the same pattern for the average precision. Next to this, also stochastic regression imputation and MICE both have on average lower precision for most levels of p.

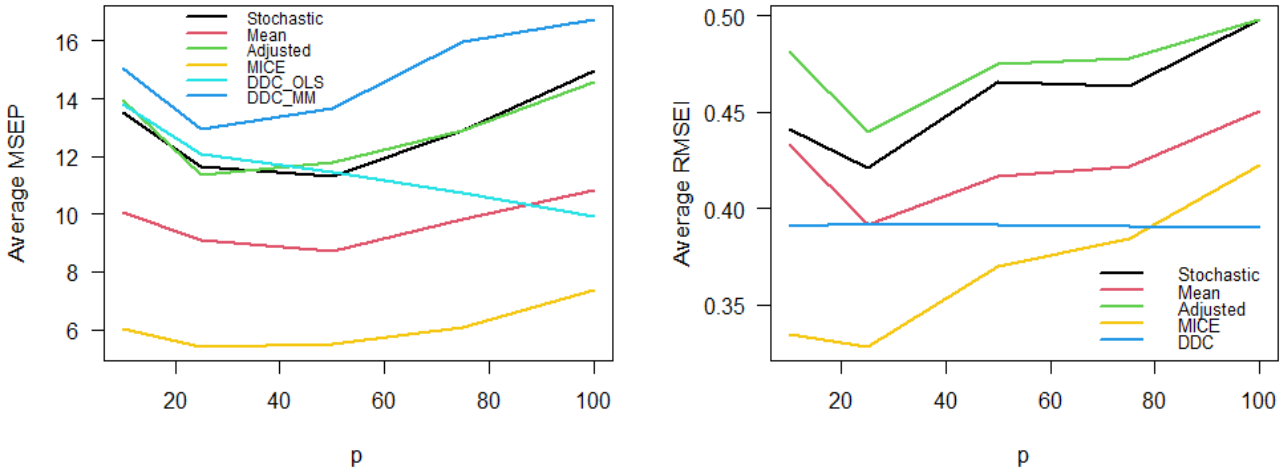

Figure 12: Average MSEP (left) and RMSEI (right) of the different MVI methods for various levels of p.

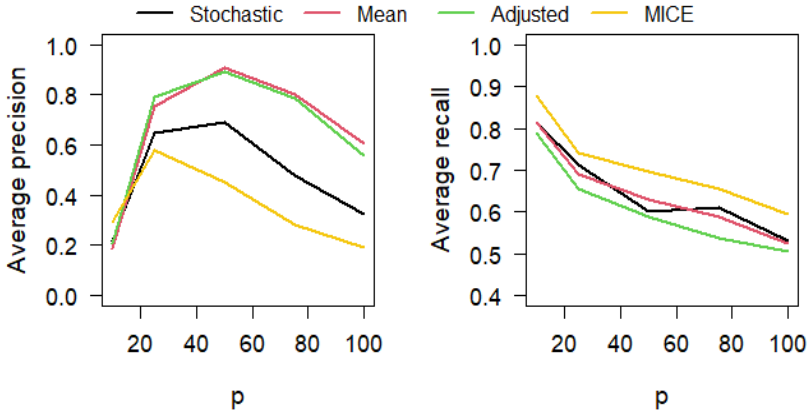

Figure 13: Average precision (left) and average recall (right) of the four imputation methods with CRM for various levels of p.

It is however, interesting to see that for all four methods, the average precision first increases with p, but then a turning point occurs and the precision decreases again. Furthermore, MICE in all situations except $p = 10$ has the lowest average precision. In contrast to this, MICE has the highest average recall for all levels of p. The right panel of Figure 13 shows an overall decreasing average recall as p increases, and again there does not seem to be a large difference between the various methods.

5 Data application

The methods researched in Section 4 are also applied to real-life data. The data are from a 2015 Swiss database of nutrition information. The goal of the statistical analysis is to find a predictive model for the dependent variable 'cholesterol'. Included are the first 200 products, for which seven contain missing values, and six variables: cholesterol, energy_kcal, protein, water, carbohydrates and sugars. All variables are logarithmically transformed, because they are right skewed. The complete results as obtained by Filzmoser et al. (2020) are given in Appendix F.

The original CRM method (with listwise deletion) is compared to the four MVI techniques combined with CRM. DDC cannot be applied on the dataset that still contains missing values,

as the dependent variable has tiny median absolute deviation, and DDC does not work in such instances. Table 1 depicts the estimated regression coefficients by the five analysed methods with CRM. From this table, it can be seen that all methods except adjusted CRM provide very comparable estimates.

Table 1: Estimated regression coefficients of Nutrients data, using four imputation methods with CRM and CRM with listwise deletion.

Variable	Estimated CRM coefficients				
	Stochastic	Mean	Adjusted	MICE	Listwise deletion
Intercept	-32.83238	-33.34892	-4.03281	-33.24908	-33.73173
log.energy_kcal	3.66342	3.64143	-0.08264	3.70199	3.62970
log.protein	0.90958	0.85222	-0.04459	0.91640	0.98341
log.water	3.61656	3.68460	-0.05031	3.65946	3.78561
log.carbohydrates	-0.09758	0.03311	0.11684	-0.16275	0.05336
log.sugars	0.05961	-0.10931	-0.05570	0.13120	-0.10999

The 10% trimmed root mean squared error of prediction (RMSEP) is compared for the five methods, after performing a 10-fold cross-validation, for which the results are in Figure 14. The RMSEP is trimmed by 10% by removing the smallest and largest 10% of prediction errors, which provides a more robust model performance evaluation by mitigating the influence of outliers. Three MVI techniques (MICE, stochastic regression and mean imputation) perform better in terms of the 10% trimmed RMSEP than listwise deletion, while adjusted CRM performs noticeably worse than all other methods. Although these three MVI methods perform quite similarly, MICE with CRM provides the most accurate predictions while also having a quite small interquartile range. Thus, it can be concluded that like in previous literature listwise deletion should not be used and even single imputation techniques can provide better estimates.

Figure 15 gives the imputed heatmap of the MICE imputation with CRM. It has detected 26 observations as casewise outliers. Each cell that is coloured (either blue or red) represents an outlying cell. If a cell was deviating upwards it is coloured red, while a cell that is coloured blue was deviating downwards. The darker the colour of a cell the more outlying it was, and therefore the bigger difference between the original and imputed value.

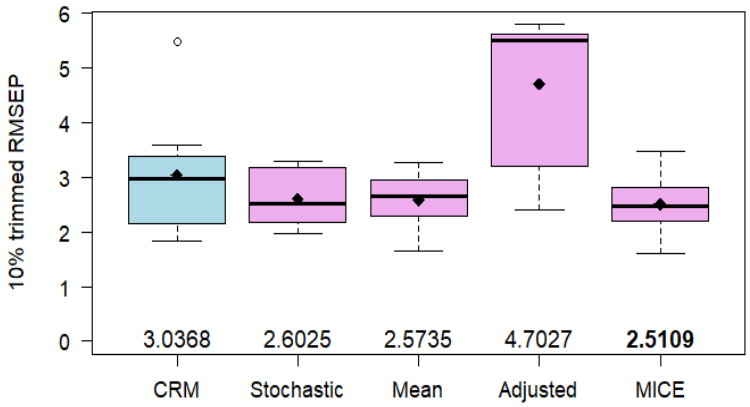

Figure 14: Boxplot of 10% trimmed RMSEP of CRM with listwise deletion and four imputation methods with CRM.

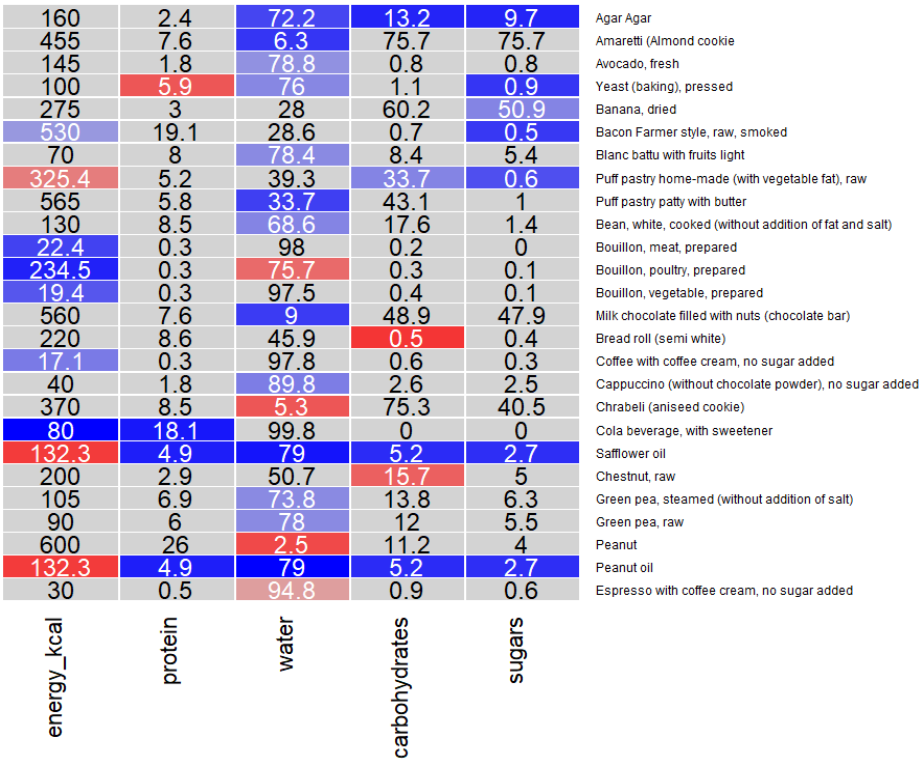

Figure 15: Heatmap of imputed outliers of MICE with CRM.

6 Conclusion and discussion

In this paper, the performance of six missing value imputation (MVI) techniques has been researched, next to reproducing the results obtained by Filzmoser et al. (2020). It is found that cellwise robust M regression (CRM) outperforms the other (robust) regression methods analysed. Furthermore, the precision and recall of CRM on average increase substantially as the cells become more outlying. In contrast, with more explanatory variables the average precision and recall of CRM decrease, so CRM becomes less precise in its detection of cellwise outliers.

The six MVI techniques researched in this paper, are stochastic regression imputation with CRM, mean imputation with CRM, an adjusted CRM method, multiple imputation by chained equations (MICE) with CRM and Detect Deviating Cells (DDC) with OLS and MM. When data are missing completely at random (MCAR), MICE with CRM overall seems to perform best in terms of limiting bias, correctly imputing the data and predictive performance. The same holds for data that are missing at random (MAR) and not missing at random (NMAR), except for the mean absolute error (MAE) of MAR data, where DDC-OLS is on average slightly better than MICE. Furthermore, analysing the average precision and recall of the four methods with CRM, it can be seen that MICE with CRM provides the best recall, but the worst precision. In contrast to this, mean imputation with CRM and adjusted CRM both have high precision but low recall, while stochastic regression imputation with CRM has the best balance. Therefore, depending on the goal (maximize precision or recall), a decision on the method can be made.

All in all, this leads to the conclusion that in most cases MICE with CRM leads to the best model performance when data are MCAR, MAR and NMAR. This is in line with the hypotheses presented earlier, where MI was expected to outperform single imputation methods (Donders et

al., 2006; Kang, 2013; Newman, 2014; Wulff & Jeppesen, 2017). However, when the goal is to maximize one specific metric, such as the precision, other methods might be better. Moreover, MICE with CRM in most cases has the slowest running time, which must be taken into account when dealing with larger datasets or when a large part of the data is missing. A good alternative in these situations is mean imputation with CRM, because it has only a slightly worse model performance but higher efficiency.

Additionally, when the percentage of missing values increases from 10% to 50%, all six evaluated methods produce more biased estimates, have worse prediction performance and less accurate imputed datasets. This trend between methods was expected, as Jadhav et al. (2019) and Madley-Dowd et al. (2019) both found no significant differences between the percentage of missing values. Furthermore, for all methods that utilise CRM, the average precision and recall decrease as the percentage of missing values increases. This implies that with more missing values, CRM becomes less able to correctly identify cellwise outliers. Next to this, with explanatory variables ranging from 10 to 100, MICE with CRM produces increasingly more biased estimates, while the other methods are not affected. Moreover, the prediction performance of MICE with CRM remains lowest for all levels analysed here, while the quality of the imputed data matrix is almost always best for MICE with CRM. This is in contrast to Kang (2013) who found that with large datasets multiple imputation techniques perform well.

In conclusion, in most cases, MICE with CRM is recommended. It produces the least biased estimates and shows the best prediction performance in almost all of the scenarios. MICE with CRM, however, does come with a major drawback. When the dataset becomes large, or contains many missing values, MICE becomes computationally expensive, which was already predicted by Azur et al. (2011). In these cases mean imputation might be more efficient, while model performance is only sacrificed somewhat.

One of the main limitations of this research is the small selection of MVI methods that were analysed. There exists a much wider variety of MVI techniques, such as Expectation-Maximization or machine learning (ML) approaches. Some commonly utilised ML methods that should be researched in combination with CRM are: k-nearest neighbours, random forests, decision trees and clustering (Lin & Tsai, 2020; Petrazzini, Naya, Lopez-Bello, Vazquez & Spangenberg, 2021). Analysing these ML imputation methods is especially important, because Hasan et al. (2021) found that statistical strategies are most commonly used, indicating ML methods are under-researched. Another limitation is that all MVI techniques were analysed when there was a fixed percentage of outliers. In the future, it would be useful to analyse how these MVI methods in combination with CRM perform as the percentage of outliers is larger. Moreover, especially when either the percentage of missing values or number of explanatory variables are varied, no single method outperforms the others on all criteria. Therefore, no specific recommendation can be made for these situations and more research is needed to find exactly which method has the best performance in all situations. Lastly, MICE with CRM is recommended in many scenarios, however, this method comes with the drawback that it can be highly computationally expensive. It would be valuable to see how the running time of MICE with CRM can be improved, such that this method can be more reasonably applied and no trade-off between performance and efficiency is necessary anymore.

References

Acock, A. C. (2005). Working with missing values. *Journal of Marriage and family*, *67*(4), 1012–1028.

Azur, M. J., Stuart, E. A., Frangakis, C. & Leaf, P. J. (2011). Multiple imputation by chained equations: what is it and how does it work? *International journal of methods in psychiatric research*, *20*(1), 40–49.

Beretta, L. & Santaniello, A. (2016). Nearest neighbor imputation algorithms: a critical evaluation. *BMC medical informatics and decision making*, *16*(3), 197–208.

Debruyne, M., Höppner, S., Serneels, S. & Verdonck, T. (2019). Outlyingness: Which variables contribute most? *Statistics and Computing*, *29*, 707–723.

De Souto, M. C., Jaskowiak, P. A. & Costa, I. G. (2015). Impact of missing data imputation methods on gene expression clustering and classification. *BMC bioinformatics*, *16*(1), 1–9.

Donders, A. R. T., Van Der Heijden, G. J., Stijnen, T. & Moons, K. G. (2006). A gentle introduction to imputation of missing values. *Journal of clinical epidemiology*, *59*(10), 1087–1091.

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

Graham, J. W. (2009). Missing data analysis: Making it work in the real world. *Annual review of psychology*, *60*, 549–576.

Graham, J. W., Olchowski, A. E. & Gilreath, T. D. (2007). How many imputations are really needed? some practical clarifications of multiple imputation theory. *Prevention science*, *8*, 206–213.

Groenwold, R. H. & Dekkers, O. M. (2020). Missing data: the impact of what is not there. *European Journal of Endocrinology*, *183*(4), E7–E9.

Hasan, M. K., Alam, M. A., Roy, S., Dutta, A., Jawad, M. T. & Das, S. (2021). Missing value imputation affects the performance of machine learning: A review and analysis of the literature (2010–2021). *Informatics in Medicine Unlocked*, *27*, 100799.

Hegde, H., Shimpi, N., Panny, A., Glurich, I., Christie, P. & Acharya, A. (2019). Mice vs ppca: Missing data imputation in healthcare. *Informatics in Medicine Unlocked*, *17*, 100275.

Heymans, M. & Eekhout, I. (2019). Applied missing data analysis with spss and (r) studio. *Heymans and Eekhout: Amsterdam, The Netherlands*.

Jadhav, A., Pramod, D. & Ramanathan, K. (2019). Comparison of performance of data imputation methods for numeric dataset. *Applied Artificial Intelligence*, *33*(10), 913–933.

Kang, H. (2013). The prevention and handling of the missing data. *Korean journal of anesthesiology*, *64*(5), 402–406.

Lin, W.-C. & Tsai, C.-F. (2020). Missing value imputation: a review and analysis of the literature (2006–2017). *Artificial Intelligence Review*, *53*, 1487–1509.

Little, R. J. & Rubin, D. B. (2019). *Statistical analysis with missing data* (Vol. 793). John Wiley & Sons.

Luengo, J., García, S. & Herrera, F. (2012). On the choice of the best imputation methods for missing values considering three groups of classification methods. *Knowledge and information systems*, *32*, 77–108.

Maddala, G. S. & Lahiri, K. (1992). *Introduction to econometrics* (Vol. 2). Macmillan New York.

Madley-Dowd, P., Hughes, R., Tilling, K. & Heron, J. (2019). The proportion of missing data should not be used to guide decisions on multiple imputation. *Journal of clinical epidemiology*, *110*, 63–73.

Newgard, C. D. & Lewis, R. J. (2015). Missing data: how to best account for what is not known. *Jama*, *314*(9), 940–941.

Newman, D. A. (2014). Missing data: Five practical guidelines. *Organizational Research Methods*, *17*(4), 372–411.

Osborne, J. W. & Overbay, A. (2004). The power of outliers (and why researchers should always check for them). *Practical Assessment, Research, and Evaluation*, *9*(1), 6.

Petrazzini, B. O., Naya, H., Lopez-Bello, F., Vazquez, G. & Spangenberg, L. (2021). Evaluation of different approaches for missing data imputation on features associated to genomic data. *BioData mining*, *14*(1), 1–13.

Pham, T. M., Pandis, N. & White, I. R. (2022). Missing data, part 1. why missing data are a problem. *American journal of orthodontics and dentofacial orthopedics*, *161*(6), 888–889.

Rousseeuw, P. J. & Bossche, W. V. D. (2018). Detecting deviating data cells. *Technometrics*, *60*(2), 135–145.

Rubin, D. B. (1976). Inference and missing data. *Biometrika*, *63*(3), 581–592.

Rubin, D. B. (2004). *Multiple imputation for nonresponse in surveys* (Vol. 81). John Wiley & Sons.

Sadouk, L., Gadi, T. & Essoufi, E. H. (2020). Robust loss function for deep learning regression with outliers. In *Embedded systems and artificial intelligence: Proceedings of esai 2019, fez, morocco* (pp. 359–368). Springer.

Salgado, C. M., Azevedo, C., Proença, H. & Vieira, S. M. (2019). *Missing data*. Springer, Cham (CH).

Scheffer, J. (2002). Dealing with missing data. *Research Letters in the Information and Mathematical Sciences*, *3*, 153–160.

Sinharay, S., Stern, H. S. & Russell, D. (2001). The use of multiple imputation for the analysis of missing data. *Psychological methods*, *6*(4), 317–329.

Van Buuren, S. (2018). *Flexible imputation of missing data*. CRC press.

Van Buuren, S., Brand, J. P., Groothuis-Oudshoorn, C. G. & Rubin, D. B. (2006). Fully conditional specification in multivariate imputation. *Journal of statistical computation and simulation*, *76*(12), 1049–1064.

Van Buuren, S. & Groothuis-Oudshoorn, K. (2011). mice: Multivariate imputation by chained equations in r. *Journal of statistical software*, *45*, 1–67.

White, I. R., Royston, P. & Wood, A. M. (2011). Multiple imputation using chained equations: issues and guidance for practice. *Statistics in medicine*, *30*(4), 377–399.

Wulff, J. N. & Jeppesen, L. E. (2017). Multiple imputation by chained equations in praxis: guidelines and review. *Electronic Journal of Business Research Methods*, *15*(1), 41–56.

Zhang, Z. (2016). Missing data imputation: focusing on single imputation. *Annals of translational medicine*, *4*(1).

A Performance of CRM without missing values and p = 50

As mentioned earlier, the following results are exactly the same as obtained by Filzmoser et al. (2020). The simulation settings are as described in Section 4.1, but now the number of explanatory variables is fixed at $p = 50$. In Figure 16 the MAE (left) and MSEP (right) for the five regression techniques (CRM, MM, DDC-MM, OLS and DDC-OLS) are given. For both the MAE and MSEP, CRM has the lowest average values, although MM regression only performs somewhat worse than CRM. Moreover, it does not seem that adding DDC as an additional step before MM regression improves bias or prediction performance. Finally, as expected the non-robust OLS regression results in the most biased regression estimates. In Figure 17 both the RMSEI of CRM and DDC (left) and the average precision and recall of CRM (right) are given. The RMSEI of CRM is lower than that of DDC, implying that CRM results on average in more accurate imputations. The recall of CRM indicates that most of the cellwise outliers have been detected as such, while the precision signifies that CRM in quite some cases incorrectly flags cells as cellwise outliers and consequently impute too many cells.

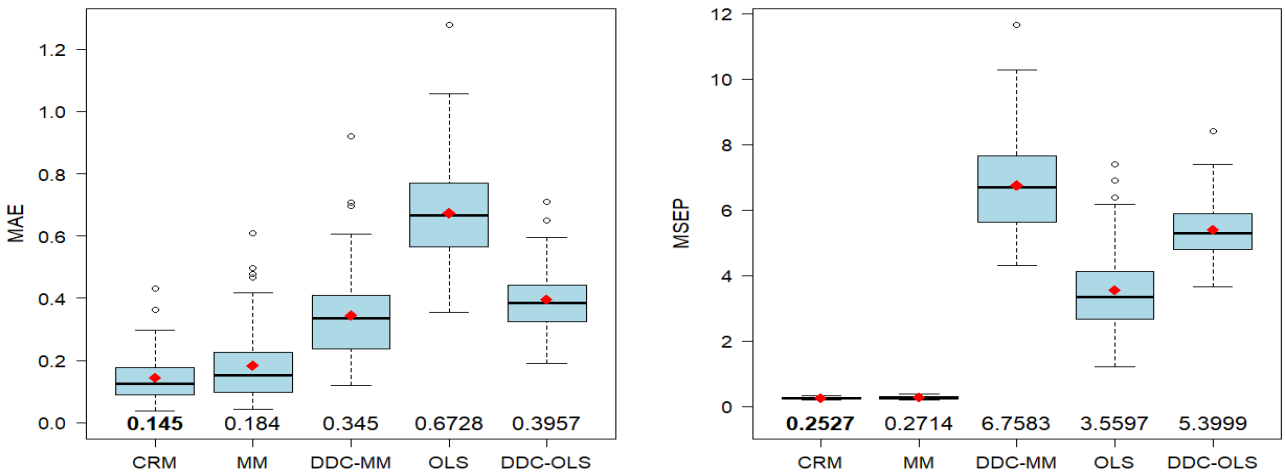

Figure 16: Boxplot of MAE (left) and MSEP (right) for the five regression methods.

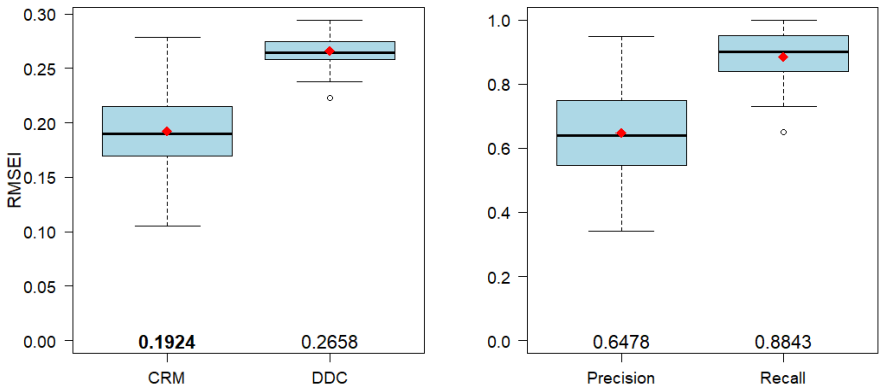

Figure 17: Boxplot of RMSEI for CRM and DDC (left) and precision and recall of CRM (right).

The degree of outlyingness, denoted by k, is varied from 0 to 8, for which the results are given in Figure 18. In this case, the simulation is repeated 10 times for each value of k and the

MAE, MSEP and RMSEI are averaged. It can be seen that the average MSEP (centre) for CRM, DDC-MM, MM and DDC-OLS is not highly affected by the level of outlyingness, as they remain relatively stable. However, OLS is affected, as when the cellwise outliers become relatively more extreme, OLS regression performs worse. A similar trend can be identified for the average MAE, where OLS is affected more by k in comparison to the other methods. Furthermore, both CRM and MM have the lowest average values of MAE and MSEP across the different levels of k. Even though the average MSEP and MAE of CRM do not differ a lot when the level of outlyingness (k) is varied, the RMSEI does. CRM on average has a lower RMSEI compared to DDC, implying more truthful imputation across the different levels of k.

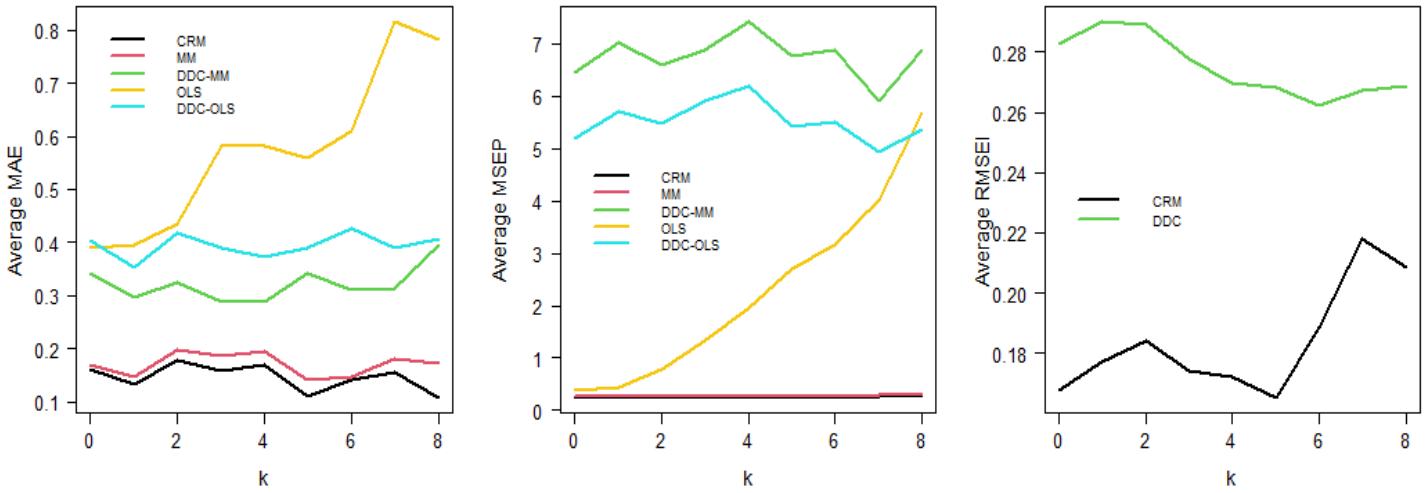

Figure 18: Average MAE (left), MSEP (centre) and RMSEI for DDC and CRM (right) for different values of k.

Figure 19 provides the average precision and recall of CRM for different levels of k. As k increases, the average precision and average recall both improve. This indicates that as the cellwise outliers become more extreme, CRM is better at identifying which cells are outliers. Moreover, as k becomes relatively large no real substantial increase can be identified for precision and recall, indicating possible diminished returns for more outlying cells.

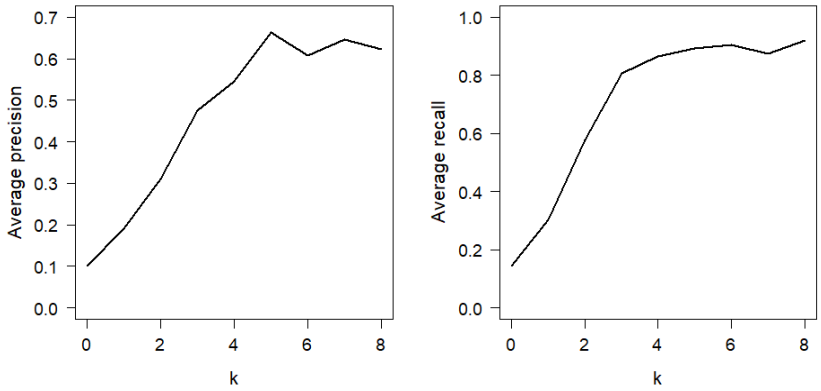

Figure 19: Average precision (left) and average recall (right) of CRM for different values of k.

Lastly, the results of the average MAE of the five regression methods with varying percentage of casewise outliers are given in Figure 20. The average MAE for both DDC-OLS and DDC-MM is relatively stable for the different levels of casewise outliers. For the remaining three methods the average MAE increases as the percentage of casewise outliers increases, most notably for OLS. This implies that for CRM, MM and OLS the bias of the regression coefficients increases with more casewise outliers, while DDC-OLS and DDC-MM are not affected in the same way.

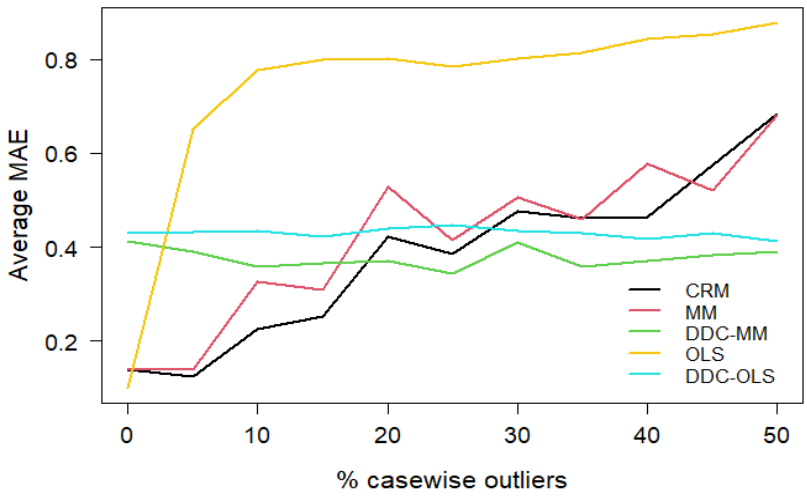

Figure 20: Average MAE for the five regression methods with different percentages of casewise outliers.

B Missing value imputation on MAR data

Instead of the data being MCAR, in this case it is assumed that the data are MAR. The most notable differences and similarities with the MCAR case are discussed in Section 4.3, while here the full results are provided. Figure 21 provides the MAE of the six MVI and regression techniques. In this figure, it can be seen that DDC-OLS instead of MICE introduces the least bias, as indicated by the lowest average MAE. Nevertheless, in Figures 22 and 23 that give respectively the MSEP and RMSEI, MICE still provides the most optimal scores. This indicates that the prediction performance and imputed matrix is still the best for MICE with CRM, like in the MCAR case. Also, looking at the RMSEI specifically, it can be seen that MICE and DDC score almost equally on their quality of imputed data matrix, while DDC has an apparent smaller interquartile range compared to MICE. The good performance of DDC on this metric is to be expected as DDC is a method designed to also handle cellwise outliers.

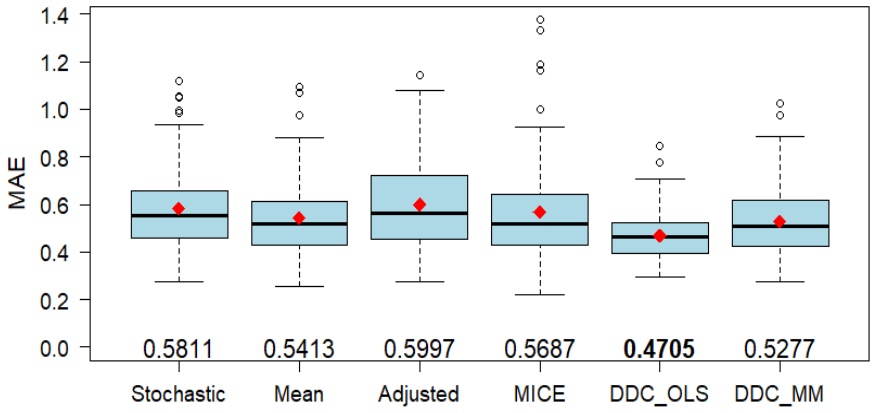

Figure 21: Boxplot of MAE for different MVI and regression methods, on MAR data.

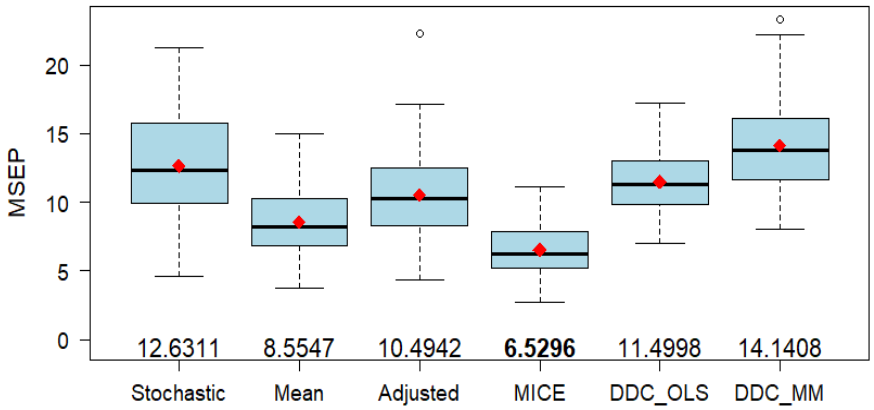

Figure 22: Boxplot of MSEP for different MVI and regression methods, on MAR data.

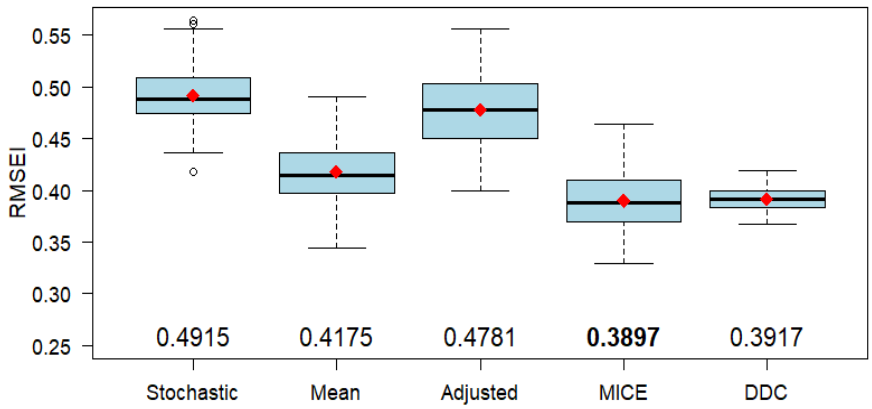

Figure 23: Boxplot of RMSEI for different imputation methods, on MAR data.

Finally, the average precision and recall for the four imputation methods with CRM are given in Figure 24. Similar to the MCAR situation it can be seen that mean imputation with CRM and adjusted CRM perform almost equally, with high average precision and moderate recall. Furthermore, stochastic regression imputation with CRM has a good balance between precision and recall, while MICE has the highest recall but lowest precision, on average. Just like in the MCAR case, making a decision on purely the precision and recall is not possible, and depends on the specific goal of the analysis.

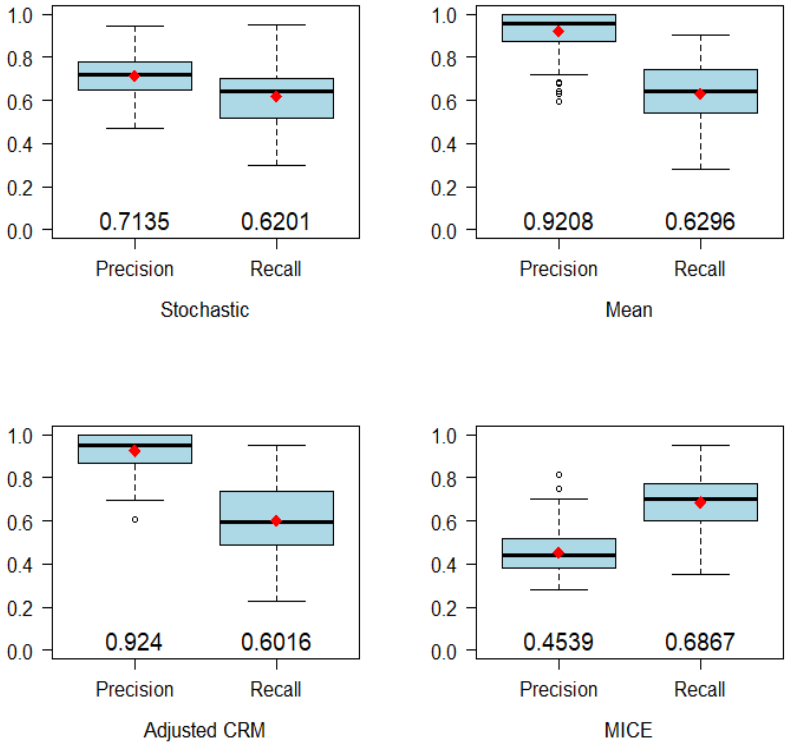

Figure 24: Boxplot of precision and recall across the four imputation methods with CRM, on MAR data.

C Missing value imputation on NMAR data

Instead of the data being MCAR or MAR, in this case it is assumed that the data are NMAR. The most notable differences and similarities with the MCAR and MAR cases are again discussed in Section 4.3, while the full results are provided here. Figure 25 depicts the MAE, where again MICE with CRM performs the best in terms of limiting bias. Although MICE has the lowest average MAE across the 100 simulation repeats, most other methods (except for Adjusted CRM) perform quite comparably. These five methods all have a MAE between 0.5 and 0.6, with relatively large and partially overlapping interquartile ranges. Therefore, purely based on MAE a real conclusion on the best performing method cannot be made.

Following this, Figure 26 displays the MSEP, which shows a clear advantage of MICE with CRM over the other five methods. The MSEP indicates that the best predictive performance is of MICE with CRM when data are NMAR, while the other methods have noticeably larger MSEP values. Also, the interquartile range of MICE is relatively small compared to the other methods, especially adjusted CRM and DDC-MM. Moreover, mean imputation which is generally regarded as a method to not be used on MAR or NMAR data, still performs remarkably well. This likely indicates that CRM acts as a buffer for the poor performance of mean imputation.

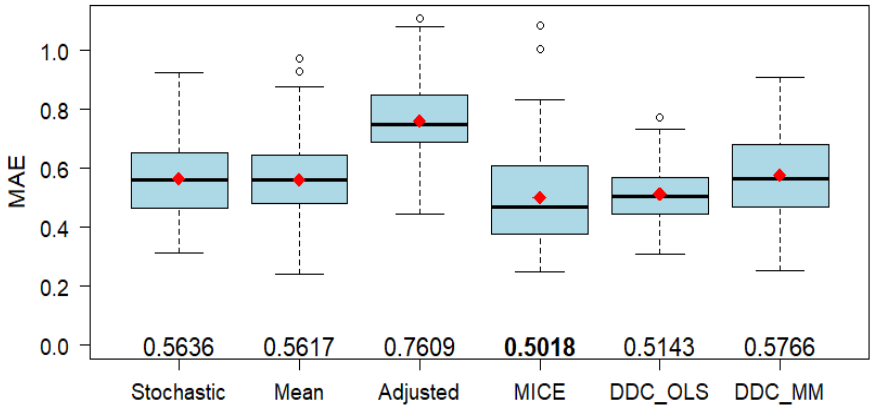

Figure 25: Boxplot of MAE for different MVI and regression methods, on NMAR data.

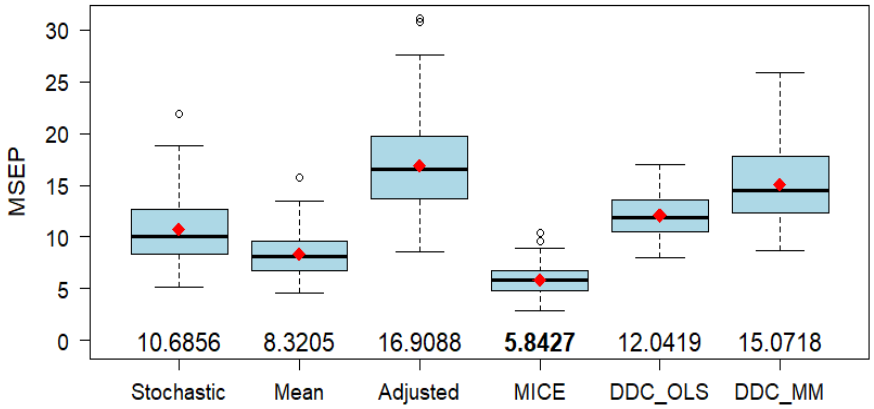

Figure 26: Boxplot of MSEP for different MVI and regression methods, on NMAR data.

Figure 27 gives the average RMSEI. Just as before, MICE imputation results in the most accurate data imputation as compared to the real dataset, but is closely followed by both DDC and mean imputation. In contrast to this, the average RMSEI of adjusted CRM is distinctly higher than the other four methods. This is expected as the missing values are only temporarily imputed in this method by 10 times the mean, which logically results in a poorly imputed dataset.

Lastly, the results of the precision and recall of the four CRM and imputation methods are depicted in Figure 28. Similar trends as with both the MCAR and MAR cases are detected.

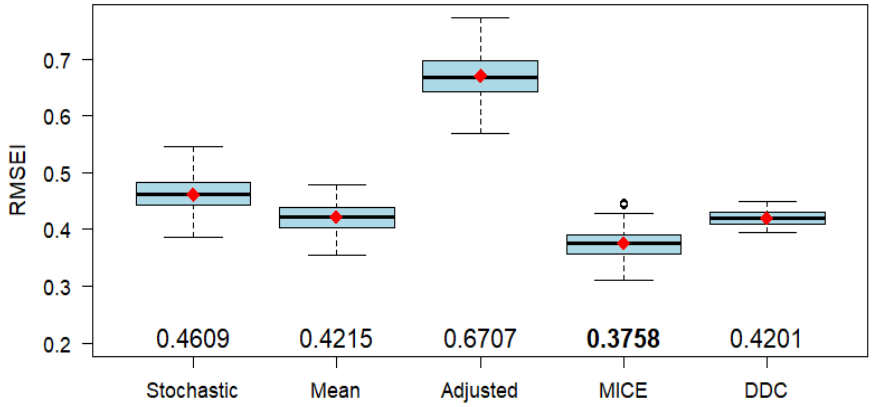

Figure 27: Boxplot of RMSEI for different imputation methods, on NMAR data.

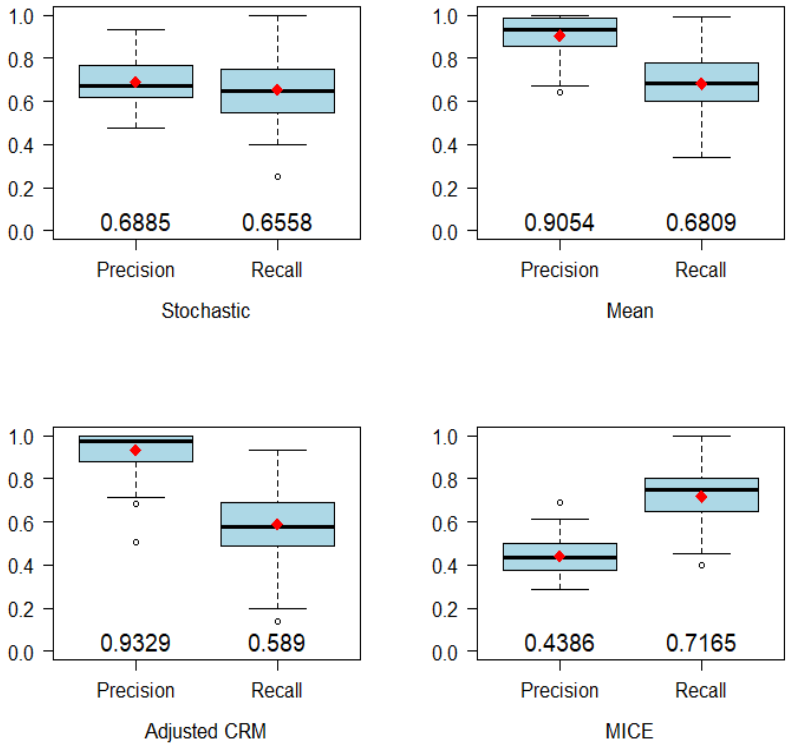

Figure 28: Boxplot of precision and recall across the four imputation methods with CRM, on NMAR data.

D Different percentage of missing values

In this section the other results obtained from analysing the performance of the imputation methods across different percentage of missing values, as performed in Section 4.4 are depicted. In Figure 29 an overall linearly increasing trend can be detected for the average MSEP as could also be seen for the average MAE in Figure 8. However, the differences that can be detected for average MAE are not necessarily present in this case. First, like for the MAE, the three methods (stochastic, mean and adjusted) behave similarly. However, in this case they perform worse than the other three methods. Moreover, the average MSEP for MICE seems to increase more linearly as compared to the situation of the average MAE. A similar trend for the average MSEP can be observed for the average RMSEI, see Figure 30. All methods show an increased average RMSEI, as the percentage of missing values increases. Moreover, both stochastic regression imputation and adjusted CRM behave worse as p increases.

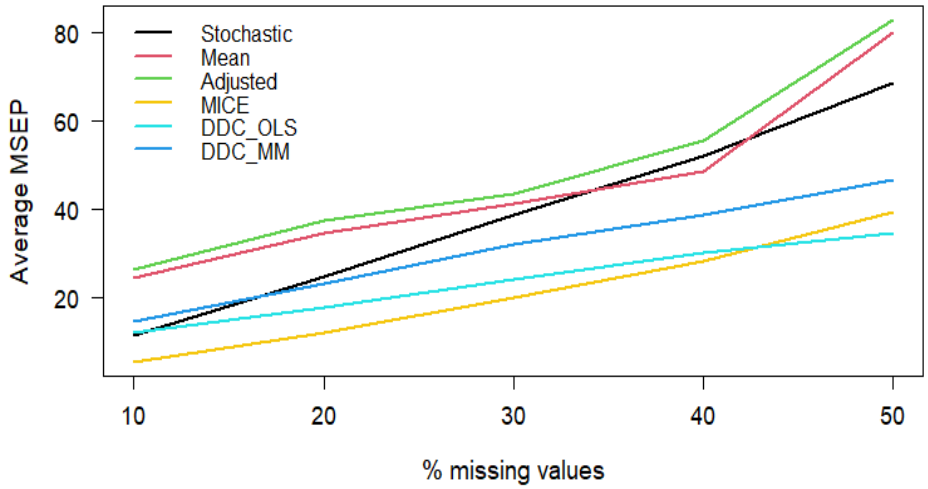

Figure 29: Average MSEP for the different MVI methods for a different percentage of missing values.

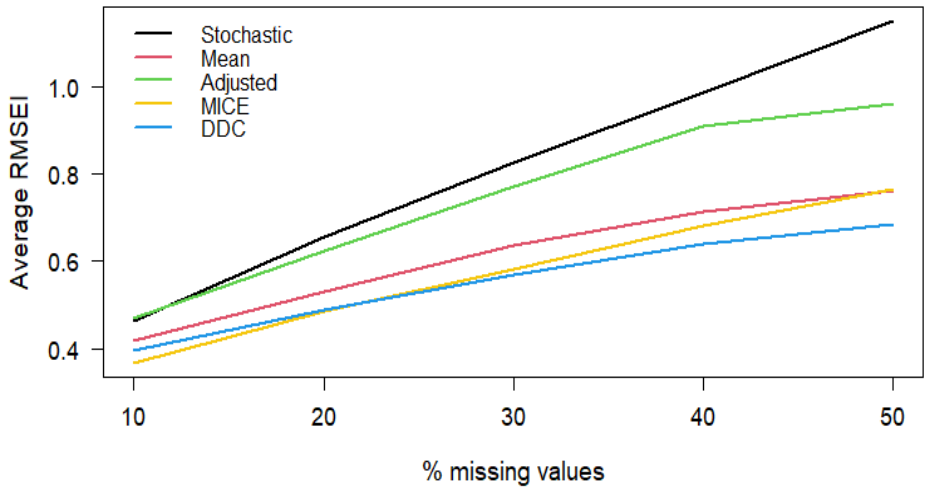

Figure 30: Average RMSEI for the different MVI methods for a different percentage of missing values.

E Comparing the performance of two MICE variations

Figure 31 shows the average MAE, MSEP and RMSEI of the two variations in MICE with CRM. One uses the percentage of missing values as an indicator of the number of datasets in MICE, while the other has a fixed number of datasets, namely 5. Generally, what can be observed is a very comparable performance of the two variations. Especially for the average MSEP (centre) and average RMSEI (right), the two MICE variations have a similar (linear) increasing trend, with the variation $m = 5$ having slightly higher values, indicating slightly worse prediction performance and less accurate imputed dataset. However, for the average MAE in the left panel this is somewhat different. Here, the $m = 5$ variation of MICE has slightly lower average values implying less biased estimates compared to the $m = \%$ variation. All in all, this indicates that the benefits of having additional datasets produced by MICE, which is what m determines, does not substantially affect model performance if it is followed by CRM. This is in contrast to existing literature, where it is often recommended to set m higher than 5 in order to limit the loss of power and provide more accurate imputations (Azur et al., 2011; White et al., 2011).

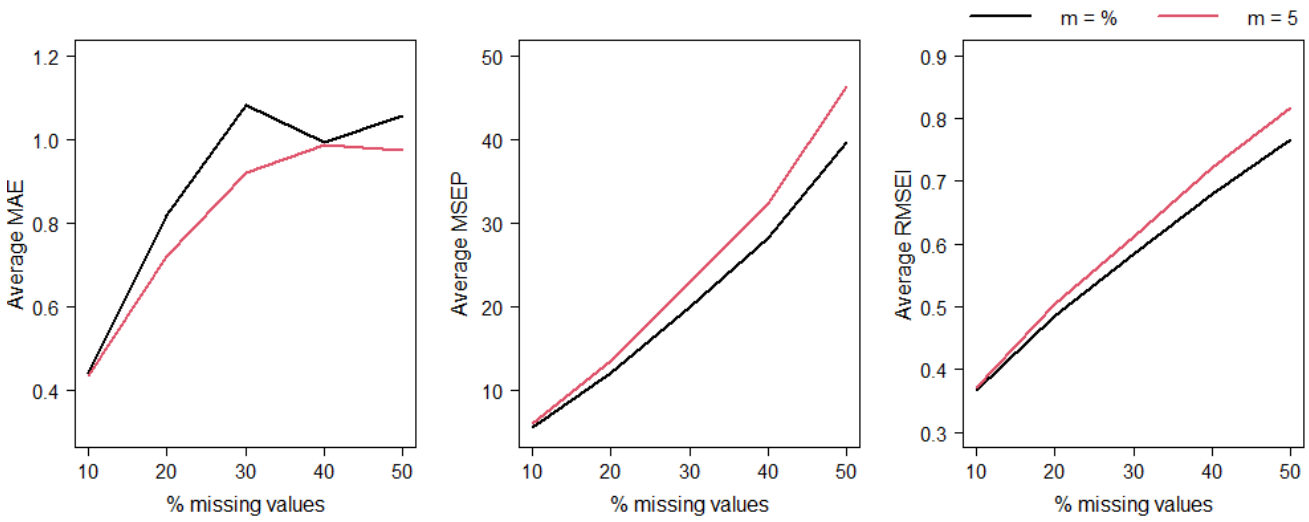

Figure 31: Average MAE (left), MSEP (centre) and RMSEI (right) of MICE with CRM for two variations.

F Data application without missing values

In this section the replication of the results of the data application by Filzmoser et al. (2020) are depicted, meaning without missing values so the method of CRM is combined with listwise deletion. Figure 32 shows the observations which were detected as casewise outliers. The coloured cells are deemed to be cellwise outliers. The darker a coloured cell the more outlying it is. It can be seen in this figure that some observations have only one outlying cell, while for others four variables are deemed outlying. Figure 33 gives the heatmap of the same observations and variables as in Figure 32, however, the outliers have now been imputed by CRM. Lastly, Figure 34 compares the five different regression techniques on the 10% trimmed RMSEP criterion. Before this RMSEP is calculated, a 10-fold cross validation is performed. As can be seen CRM has the lowest 10% trimmed RMSEP, indicating the most accurate predictions, likely because it can make use of more uncontaminated information. Nevertheless, the differences between CRM and both OLS and DDC-OLS are not substantial, with both OLS and DDC-OLS having a smaller interquartile range.

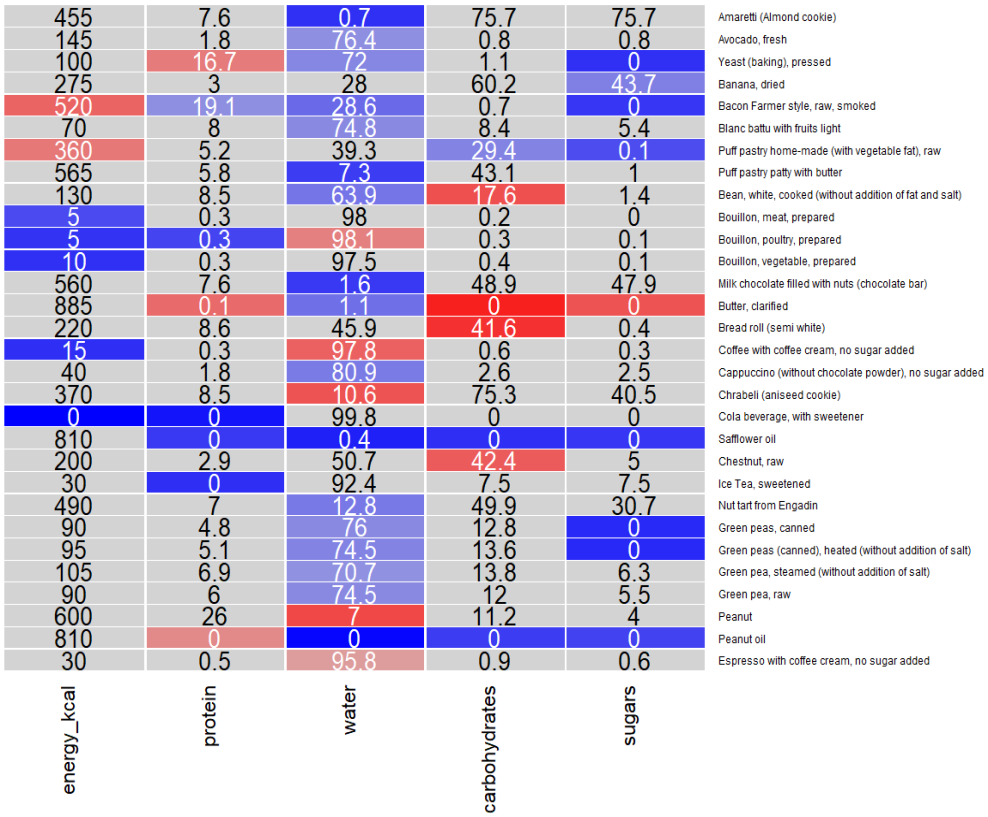

Figure 32: Heatmap of detected outliers by CRM.

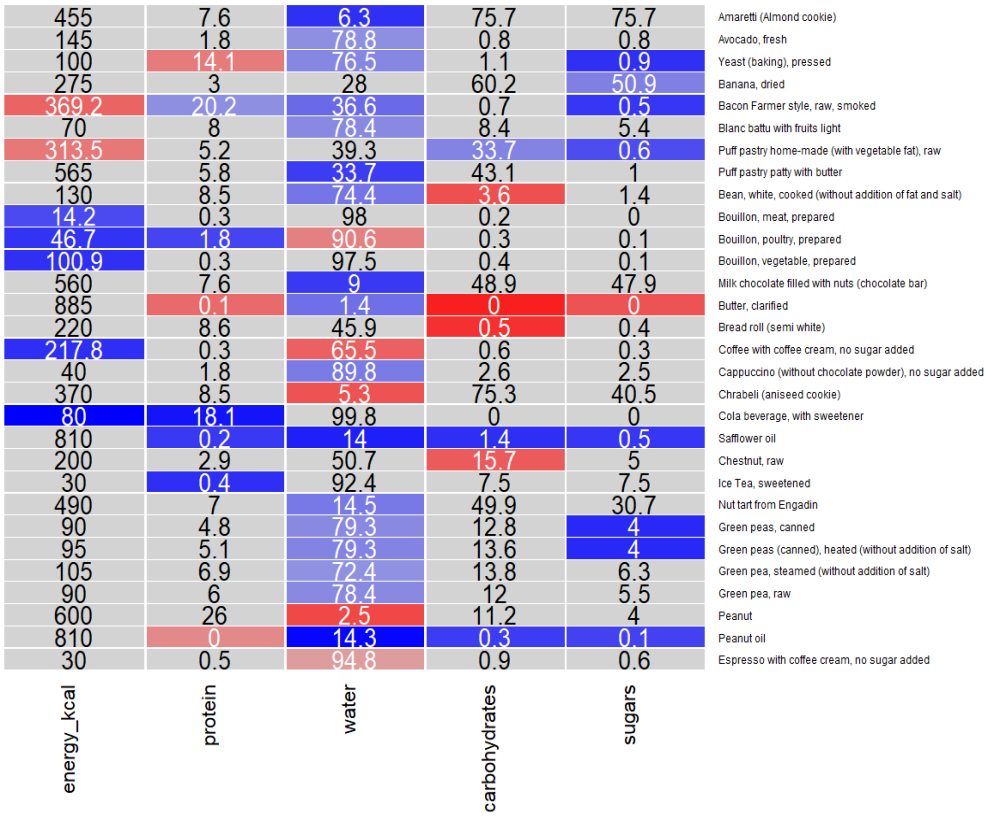

Figure 33: Heatmap of imputed outliers by CRM.

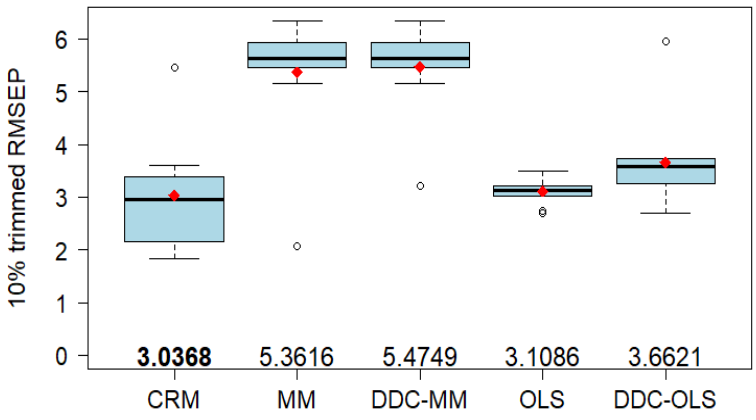

Figure 34: Boxplot of 10% trimmed RMSEP for the five regression methods.

G README file: Missing value imputation and CRM

G.1 Code description

The R scripts in this repository can be used to reproduce the results for the Bachelor thesis entitled: Enhancing prediction performance in the context of Cellwise Robust M Regression: a comparison of missing value imputation techniques (Esmée Mulder, 523280).

The code is based on the simulation study performed by Filzmoser, P., Höppner, S., Ortner, I., Serneels, S., and Verdonck, T. (2020) in their paper entitled: Cellwise robust M regression. The simulation study has been extended to analyse the performance of different missing value imputation techniques in combination with Cellwise robust M regression (CRM). The code for the original simulation study by Filzmoser et al. (2020) as well as the `crmReg` package is available using the following GitHub link: `https://github.com/SebastiaanHoppner/CRM`.

G.2 Code content

This folder consists of the following R scripts, with the first three scripts reproducing the results obtained by Filzmoser et al. (2020):

- `SimulationStudyCRMBaseline.R`: this R script performs the baseline simulation study comparing CRM to other (robust) regression techniques. It compares the methods on certain evaluation criteria and produces subsequent figures.
- `SimulationStudyCRMOutlyingness.R`: a similar simulation study is performed, but the level of outlyingness of the cellwise outliers is varied. Evaluation criteria are calculated and appropriate figures are produced.
- `SimulationStudyCRMCasewiseOutliers.R`: a similar simulation study, but with varying percentages of casewise outliers. Again evaluation criteria and their figures are produced.
- `OriginalDifferentNumberofP.R`: this R script extends the original simulation study by analysing the performance of CRM against the other regression techniques on a different number of explanatory variables. Appropriate figures are also produced in this script.
- `MissingValueImputation.R`: next to outliers, missing values are introduced. This R script compares 6 different missing value imputation and regression techniques and analyses them across the different evaluation criteria and produces the necessary figures.
- `DifferentPercentageMissingValues.R`: the effect on the performance of the 6 evaluated methods is analysed in this R script, with a varying percentage of missing values. Next to producing the appropriate figures, a comparison between two MICE variations is performed.
- `DifferentNumberOfExplanatoryVariables.R`: this R script analyses the effect of a different number of explanatory variables on the 6 imputation and regression techniques.
- `DataApplication.R`: performs a data application with and without missing values, also reproducing part of the results obtained by Filzmoser et al. (2020).