RWTH Aachen University

Institut für Statistik und Wirtschaftsmathematik

— Prof. Dr. M. Kateri —

Seminar zur Statistik und stochastischen Modellierung

Wintersemester 2022/23

Gedeon Alexander Vogt

The algorithmic Search for the optimal Number of Imputations

27. März 2023

Contents

1 Preface

The subject of study of this paper is a method that aims to tackle a widespread problem: missing data. One possible solution to this problem is the so-called multiple imputation (MI) method, which can be briefly summarized as imputing multiple values for each missing data point. This raises the question, what the optimal number of imputation would be.

The objective of this work is to give an answer to this question by reviewing already existing approaches and a relatively new one, which is an iterative procedure that was developed by Nassiri et al. in 2020. In the second part of the paper we will examine this algorithm in more detail and put it to the test with the help of a simulation study.

The outline is as follows: Section 2 will give an introduction to the general idea of multiple imputation followed by the derivation of the theoretical background in Section 3, which will be used in the end of this section to give a reasoning for the determination methods in the literature. After that, in Section 4, Nassiri et al.'s algorithm will be introduced and it will be concluded with the above mentioned simulation study, which will demonstrate a decent performance for a variety of settings in Section 5.

2 Multiple Imputation: An Introduction

Dealing with incomplete data can be a struggle, since methods like calculating a mean or fitting a regression cannot be applied to data sets with missing data points. However, there are ways to fix that. One solution could be to simply delete incomplete observations. In some cases this would be a sufficient solution, but in some cases, this could lead to inaccurate inferences, since concerns can be raised whether the incomplete observations differ systematically from the complete ones (Rubin 1978, p. 20).

Another way would be the so-called Single Imputation, where for each missing datum a single, potentially fitting value is imputed (Rubin, 1987, p. 11). This has the advantage of not losing any information by dropping incomplete rows of the data matrix and still being able to use standard complete-data methods. An example to determine the values that are to be imputed can be a regression, which gives an estimate of the missing value based on the other features of the observation.

The method, that is subject to this seminar paper is similar to single imputation with the differences, that not only one, but $k \geq 2$ imputations are made for each missing value which represent a distribution of suitable values (Rubin and Schenker 1986, p. 366). This leads to k different, complete data sets, which have to be analyzed separately and the results combined to a single estimator. This procedure is called multiple imputation and was initially developed by Rubin in his seminal work in 1978.

In multiple imputation, two or more values are drawn from the posterior distribution of missing data (the distribution that conditions on the observed data). Thus a great way to imagine this procedure is by seeing multiple imputation as a simulation.

Rubin (1987, p. 3) states, that multiple imputation (MI) is particularly beneficial in the context of survey analysis. One reason he gives is, that survey data, for example collected by the Census Bureau, are often meant to be published and made accessible to a broader number of people. Hence, it might be from interest, that the data sets are complete and the imputation process is somehow comprehensible. Second, missing values in surveys generally do not occur randomly. This means, that the distribution of missing values might differ systematically from the complete ones and assuming a distribution for those values is in general rather wrong than right.

Furthermore, there are more advantages and some disadvantages which are examined in the following section.

2.1 Advantages and Disadvantages of MI

We start with the most obvious advantage of imputation methods, that standard complete-data procedures are applicable again, no matter whether using single imputation or multiple imputation (Rubin, 1987, p. 15).

However, the major drawback of single imputation is, that during the analysis of imputed data sets, the imputed values are treated as if they were known (Rubin and Schenker, 1986, p. 366). The issue with that is, that in general the single value we are about to impute, cannot be correct (Rubin 1978, p. 21). This means, that by treating the imputed values as if they were the real ones, the actual variance of our estimators is generally underestimated, since the extra amount of uncertainty (i.e. variance) regarding the imputed value is not captured (Rubin and Schenker, 1986, p. 366). Hence, analyzing multiple data sets with different reasonable imputed values (we can also assume different models and distributions from which we draw the imputation values for each of the D data sets), is the 'most natural way to display [...] sensitivity' (Rubin, 1978, p. 21). In the case that the imputations are from multiple models, the uncertainty regarding the correct model is also properly displayed by how much the inferences differ from model to model (Rubin and Little, 2002, p. 85). This, the sampling variability and additional uncertainty, is impossible to reflect by single imputation, regarding to Rubin (1987, p. 11).

This is complemented by the fact, that MI is quite easy to implement (Carpenter and Kenward, 2008, p. 77). As Rubin (1978, p. 20) suggests, you only have to replace the missing values with a pointer to a D-dimensional vector that contains the imputation values. The procedure of analysis will be the same for all D data sets. So it can be programmed with a ordinary loop or at least with manageable number of modifications, like Rubin (1987, p. 4) notes. As we will see later in section 3.2.1, to get a valid estimator, an average over the D data set specific estimators is sufficient.

This leads us to the first disadvantage. The amount of work that goes into analyzing D data sets is D times the work needed for single imputation. In this context more work means, that the running time of the program is increased (Rubin and Little, 2002, p. 86). In smaller, less complex programs, this difference is barely noticeable, but it can have a huge impact when diving into big data and train a neural network D-times.

Another advantage is the knowledge of the data collector that goes into the process of creating appropriate imputes (Rubin and Little, 2002, p. 85). When we reconsider the example for a data collector from above, the Census Bureau, we can easily see why their imputations might be from higher quality than those of independent researchers. The Census Bureau usually has access to more information, that might be correlated to our incomplete data set and thus provides a better posterior distribution to draw from. For example zip codes of dwelling units might help to close the gaps of, for example, annual incomes (Rubin and Little, 2002, p. 85).

2.2 MI applied in Practice: An Example

Before we dive into the derivation of the mathematics behind MI and study some properties, it might be useful to give an example of an incomplete data set and how to handle it with multiple imputation.

The Leuven Eye Study (Pinto et al., 2015) was an observational study of glaucoma (an eye disease that can result in vision loss) at UZ Leuven. The aim of the study was to determine the factors that cause glaucoma. The response variable was in this case treated as binary: either the patient has glaucoma or not. A good way to identify those factors is with the help of a logistic regression. The study measured 141 characteristics for 585 patients. The risk factors include the physical condition of the eye, the previous medical history of the patients and some further biometrical measurements like blood pressure. Since this is a lot of data that has to be gathered, it was to be expected that some data would be missing in the end and thus lead to an incomplete data set. Those missing data could be caused by for example, mistakes in the measurement, the unwillingness of patients to disclose some information or simply the missing to even do some measurements, since a hospital usually a quite busy place.

However, Nassiri et al. (2020, p. 132-134) used their iterative multiple imputation procedure (imi), which will be described in more detail in section (4.2), on this incomplete data set. Since the rate of missing data is about 35% and because of the various variable types, for which different estimation techniques are required, the algorithm suggested 58 imputations until it considers the estimators of the logistic regression's parameters as converged.

3 Properties of Multiple Imputation

In order to be able to derive the basic characteristics of MI, some knowledge of Bayesian statistics is required. Therefore, we will first elaborate the necessary results.

3.1 Bayes Theorem

Let f denote a probability density and let X, Y, Z be random variables. Based on the Bayes' theorem for events (c.f. Kamps and Cramer, 2021, p. 41) the Bayes' theorem for continuous random variables can be formulated as

$$f_{X|Y=y}(x) = \frac{f_{X,Y}(x,y)}{f_Y(y)}, \quad \text{for } f_Y(y) \neq 0. \tag{3.1}$$

Thus it can be concluded, that

$$f_{Y|X=x}(y) \overset{(3.1)}{=} \frac{f_{X,Y}(x,y)}{f_X(x)} \, \frac{f_Y(y)}{{}_Y(y)} = \frac{f_{X,Y}(x,y)}{f_Y(y)} \, \frac{f_Y(y)}{f_X(x)} \overset{(3.1)}{=} f_{X|Y=y}(x) \, \frac{f_Y(y)}{f_X(x)}. \tag{3.2}$$

The Law of total probability, which is mentioned in Kamps and Cramer (2021, p. 45) and here adapted for continuously distributed random variables, is formulated as

$$f_Y(y) \;=\; \int f_{X,Y}(x,y)\mathrm{d}x. \tag{3.3}$$

So, when introducing another random variable, C, we can conclude that

$$f_{Y|Z=z}(y) \;=\; \int f_{X,Y|Z=z}(x,y)\mathrm{d}x. \tag{3.4}$$

Proof:

$$\begin{aligned}
f_{Y|Z=z}(y) &= \frac{f_{Y,Z}(y,z)}{f_Z(z)} \\
&\overset{(3.3)}{=} \frac{1}{f_Z(z)} \int f_{X,Y,Z}(x,y,z)\mathrm{d}x \\
&= \frac{1}{f(C)} \int \frac{f_{X,Y,Z}(x,y,z)}{f_Z(z)}\, f_Z(z)\mathrm{d}x \\
&= \frac{1}{f_Z(z)} \int f_{X,Y|Z=z}(x,y)\, f_Z(z)\,\mathrm{d}x \\
&= \int f_{X,Y|Z=z}(x,y)\mathrm{d}x
\end{aligned}$$

□

As the expected value is defined as

$$E_X(X) \;=\; \int x\, f_X(x)\,\mathrm{d}x, \tag{3.5}$$

we can write

$$E_X(X|Y) \overset{\text{Def. (3.5)}}{=} \int x\, f_{X|Y=y}(x)\,\mathrm{d}x \overset{(3.4)}{=} \int\int x\, f_{X,Z|Y=y}(x,z)\,\mathrm{d}x\,\mathrm{d}z. \tag{3.6}$$

Now we can phrase our first property that is used later.

3.1 Proposition: (Law of total Expectation)
Let X, Y, Z be random variables then it holds true

$$E_X(X|Y) \;=\; E_Z[E_X(X|Y,Z)|Y].$$

Proof:

$$\begin{aligned}
E_X(X|Y) &\overset{(3.6)}{=} \int\int x\, f_{X,Z|Y=y}(x,z)\,\mathrm{d}x\,\mathrm{d}z \\
&\overset{(3.1)}{=} \int\int x\, f_{X|Y=y,Z=z}(x)\, f_{Z|Y=y}(z)\,\mathrm{d}x\,\mathrm{d}z \\
&= \int\int x\, f_{X|Y=y,Z=z}(x)\,\mathrm{d}x\, f_{Z|Y=y}(z)\,\mathrm{d}z \\
&\overset{(3.6)}{=} \int E_X(X|Y,Z)\, f_{Z|Y=y}(z)\,\mathrm{d}z \\
&= E_Z[E_X(X|Y,Z)|Y]
\end{aligned}$$

□

Using the definition of the variance

$$Var_X(X) = \int x^2\, f_X(x)\,\mathrm{d}x - E_X(X)^2,$$

we can follow the following expression, which will be needed in the proof of the law of total variance below:

$$(3.7) \qquad Var_X(X|Y) \;=\; \int \left(x^2\, f_{X|Y=y}(x) - E_X(X|Y)^2\right)\,\mathrm{d}x \;=\; E_X(X^2|Y) - E_X(X|Y)^2.$$

3.2 Proposition: (Law of total Variance)

$$Var_X(X|Y) \;=\; E_Z[Var_X(X|Y,Z)|Y] + Var_Z[E_X(X|Y,Z)|Y]$$

Proof: Let X, Y, Z be random Variables. Then with (3.7) and Proposition (3.1) we can write

$$(3.8) \quad E_X(X^2|Y) \overset{\text{Prop. }(3.1)}{=} E_Z[E_X(X^2|Y,Z)|Y] \overset{(3.7)}{=} E_Z[Var_X(X|Y,Z) + E_X(X|Y,Z)^2|Y].$$

We can use this expression to conclude

$$\begin{aligned}
Var_X(X|Y) &\overset{(3.8)}{=} E_Z[Var_X(X|Y,Z) + E_X(X|Y,Z)^2|Y] - E_Z[E_X(X|Y,Z)|Y]^2 \\
&= E_Z[Var_X(X|Y,Z)|Y] + E_Z[E_X(X|Y,Z)^2|Y] - E_Z[E_X(X|Y,Z)|Y] \\
&\overset{(3.7)}{=} E_Z[Var_X(X|Y,Z)|Y] + Var_Z[E_X(X|Y,Z)|Y]
\end{aligned}$$

□

3.2 Properties

Data that is not missing at random means, that there is a cause for it. For example, censuses in the past have shown that high- and low-income households were less likely to report their income than the average household. Thus, those missing information do not occur at random but is dependent on the income. In contrast to that, if data is missing at random without any visible cause, we talk about 'ignorable mechanisms' or 'missing at random mechanisms' (Rubin, 1987, p. 3). Even though, in the event of ignorable mechanisms multiple imputation is not needed, since the available data would be sufficient to draw valid inferences (Rubin and Schenker, 1986, p. 367), we can examine this case to reveal some useful properties.

3.2.1 Expected Value and Variance

Let be ϑ the parameter we want to estimate. Rubin and Little (2002, p. 200) define the posterior distribution for a model with ignorable missing mechanisms as

$$\pi(\vartheta|Y_{obs}, D) \equiv \pi(\vartheta|Y_{obs}) = constant \times \pi(\vartheta) \times f(Y_{obs}|\vartheta),$$

with D referring to the number of imputations, $\pi(\vartheta)$ being the prior distribution and $f(Y_{obs}|\vartheta)$ the density of the observed data given ϑ, also called sample distribution. The posterior distribution is the central idea of the Bayesian approach (Robert, 2007, p. 23). It is the distribution of ϑ after the data Y_obs is observed, while the prior distribution is the assumed probability distribution before any observations are made. So, the $\pi(\vartheta|Y_{obs})$ can be seen as an updated version of ones initial beliefs. Notice, that the prior and posterior distribution may not differ from each other. This could be the case if Y_{obs} does not depend on ϑ (Robert, 2007, p. 23). With Proposition (3.1) and Proposition (3.2), we can write

$$E_{\vartheta}(\vartheta|Y_{obs}) = E_{Y_{mis}}[E_{\vartheta}(\vartheta|Y_{mis}, Y_{obs})|Y_{obs}] \quad \text{and} \tag{3.9}$$

$$Var_{\vartheta}(\vartheta|Y_{obs}) = E_{Y_{mis}}[Var_{\vartheta}(\vartheta|Y_{mis}, Y_{obs})|Y_{obs}] + Var_{Y_{mis}}[E_{\vartheta}(\vartheta|Y_{mis}, Y_{obs})|Y_{obs}]. \tag{3.10}$$

According to Rubin and Little (2002, p. 210), $\pi(\vartheta|Y_{obs})$ can be approximated as

$$\pi(\vartheta|Y_{obs}) \approx \frac{1}{D}\sum_{d=1}^{D} \pi(\vartheta|Y_{mis}^{(d)}, Y_{obs}), \tag{3.11}$$

where $Y_{mis}^{(d)} \sim \pi(Y_{mis}|Y_{obs})$ are draws of Y_{mis} from the posterior distribution that predicts missing values given the observed ones (Rubin and Little, 2002, p. 210).

3.3 Proposition: (Approximated Expected Value of ϑ)
The expected values of ϑ can be approximated as

$$E_{\vartheta}(\vartheta|Y_{obs}) \approx \bar{\vartheta},$$

where $\bar{\vartheta}$ is the corresponding MI estimator.

Proof:

$$E_\vartheta(\vartheta|Y_{obs}) \overset{(3.11)}{\approx} \int \vartheta \, \frac{1}{D} \sum_{d=1}^{D} \pi(\vartheta|Y_{mis}^{(d)}, Y_{obs}) \, \mathrm{d}\vartheta \overset{\text{not neg.}}{=} \frac{1}{D} \sum_{d=1}^{D} \int \vartheta \, \pi(\vartheta|Y_{mis}^{(d)}, Y_{obs}) \, \mathrm{d}\vartheta$$

$$= \frac{1}{D} \sum_{d=1}^{D} E_\vartheta(\vartheta|Y_{mis}^{(d)}, Y_{obs}) = \frac{1}{D} \sum_{d=1}^{D} \hat{\vartheta}_{(d)} = \bar{\vartheta}$$

Here, $\hat{\vartheta}_{(d)}$ denotes the estimate for ϑ from the dth imputed data set. □

Similarly we can approximate

$$(3.12) \qquad Var_\vartheta(\vartheta|Y_{obs}) \approx \frac{1}{D} \sum_{d=1}^{D} Var_\vartheta(\vartheta|Y_{mis}^{(d)}, Y_{obs}) + \frac{1}{D-1} \sum_{d=1}^{D} [E_\vartheta(\vartheta|Y_{mis}^{(d)}, Y_{obs}) - \bar{\vartheta}]^2.$$

The prefactor of the second summand, $1/(D-1)$ is due to the sample variance, to obtain an unbiased estimator.

3.4 Definition: (Within- and Between Variability)

The summands of (3.12) can be denoted

$$\hat{W} := \frac{1}{D} \sum_{d=1}^{D} Var_\vartheta(\vartheta|Y_{mis}^{(d)}, Y_{obs})$$

$$\hat{B} := \frac{1}{D-1} \sum_{d=1}^{D} [E_\vartheta(\vartheta|Y_{mis}^{(d)}, Y_{obs}) - \bar{\vartheta}]^2 = \frac{1}{D-1} \sum_{d=1}^{D} (\hat{\vartheta}_d - \bar{\vartheta})^2.$$

$\hat{W}$ can be seen as the average variance of $\hat{\vartheta}_{(d)}$, the variance of the estimator for ϑ over all imputed data sets. On the other hand, $\hat{B}$ is the variability of the estimator between the imputed data sets (Nassiri et al., 2020, p. 125).

Up to here, we treated ϑ as a one dimensional parameter. But the above properties also hold true when ϑ is a vector. In this case, we get

$$\hat{W} := \frac{1}{D} \sum_{d=1}^{D} \hat{\Sigma}_d$$

$$\hat{B} := \frac{1}{D-1} \sum_{d=1}^{D} (\hat{\vartheta}_{(d)} - \bar{\vartheta})(\hat{\vartheta}_{(d)} - \bar{\vartheta})',$$

according to Rubin (1987, p. 35), where $\hat{\Sigma}_d$ is the estimated covariance matrix.

3.2.2 Asymptotic Distribution and finite Imputations Correction

In this section we want to examine, how $\bar{\vartheta}$ is distributed in dependence on D. For simplification, we consider the one dimensional case of ϑ, which can however be easily transferred to the multi dimensions. Starting with the case of $D \longrightarrow \infty$, we formulate the following corollary that is directly induced by the central limit theorem.

3.5 Corollary:
Let $(\hat{\vartheta}_{(d)})_{n\in\mathbb{N}}$ be a sequence of random variables, that are i.i.d. with $\sigma^2 = Var(\hat{\vartheta}_{(1)}) = \hat{B} + \hat{W} < \infty$ and $\mu = E(\hat{\vartheta}_{(1)}) = \vartheta$. If $\hat{\vartheta}_{(d)}$ is a random vector and $\hat{\vartheta}_{(1)}, \hat{\vartheta}_{(2)}, \hat{\vartheta}_{(3)}, ...$ i.i.d., then for $D \longrightarrow \infty$, $\bar{\vartheta}$ is distributed as:

$$\bar{\vartheta} \overset{\text{as}}{\sim} N(\vartheta, \hat{B} + \hat{W}).$$

Since this holds true only asymptotically, Rubin and Schenker (1987, p. 367) suggest to add the factor $(1 + D^{-1})$, to account for the extra variability in the estimator for a finite number of imputations.

3.6 Proposition: (Total Variance for finite D)
The total variance for finite D can be written as

$$\hat{V}_D \; := \; (1 + D^{-1})\, \hat{B} + \hat{W}.$$

For $D \longrightarrow \infty$ we get $\hat{V} := \hat{B} + \hat{W}$.

Proof: This is due to the fact, that the variance of the mean of the estimators for the parameter of interest, $\bar{\vartheta}_D$, based on D imputations is

$$\begin{aligned} Var(\bar{\vartheta}_D) & \overset{(3.10)}{=} E[Var(\bar{\vartheta}_D|Y_{obs})] + Var[E(\bar{\vartheta}_D|Y_{obs})] \\ &= \hat{B}/D + \hat{V} \\ &= \hat{B}/D + \hat{B} + \hat{W} \; = \; (1 + D^{-1})\, \hat{B} + \hat{W} \end{aligned}$$

(c.f. Rubin and Schenker, 1986, p. 367). □

Furthermore, Rubin and Schenker (1986) discovered, that for small D, we can approximate with the help of the Student's t distribution, t_ν:

$$(\bar{\vartheta}_D - \vartheta) \sim \sqrt{\hat{V}_D}\, t_\nu \quad \Longleftrightarrow \quad \bar{\vartheta}_D \approx \vartheta + \sqrt{\hat{V}_D}\, t_\nu.$$

Also according to Rubin and Schenker (1986, p. 367), the degrees of freedom ν can be computed with

$$\nu = \left[1 + \left(\frac{D}{D+1}\right) \frac{\hat{W}}{\hat{B}}\right]^2 (D-1).$$

This means, that for large $\hat{W}/\hat{B}$ (this could happen due to a high response rate or few missing values) $\sqrt{\hat{V}_D}\, t_\nu$ would be close to $N(0, \hat{V}_D)$. On the other and if $\hat{W}/\hat{B}$ is small, ν would be approximately equal to $(D-1)$ (Rubin and Schenker, p. 367).

3.2.3 Information depending on the missing Data

The variance expressions we derived in Section 3.2.2 could be used to provide some insights on how much information was lost due to the missing data.

In the case that we have a complete data set with no missing data points, applying multiple imputation would produce the same data set over and over again. So the variance between the imputed data sets, $\hat{B}$, would be equal to zero. So Carpenter and Kenward (2008, p. 79) described the variance increase which is caused by the proportion of missing data as

$$\left(\frac{\hat{W}+\hat{B}}{\hat{W}}\right) 100\% \;=\; \left(1+\frac{\hat{B}}{\hat{W}}\right). \tag{3.13}$$

Since we assume an asymptotic normal distribution, we can say, that 'information' can be written as $1/variance$.

Proof: The Fisher-Information (Fisher, 1935, p. 220ff) is defined as

$$I(\vartheta) \;:=\; Var_{\vartheta}(S_{\vartheta}), \quad \text{with} \tag{3.14}$$

$$S_{\vartheta}(x) \;:=\; \frac{\partial}{\partial\vartheta}\ln f_{\vartheta}(x). \tag{3.15}$$

The function f_{ϑ} is a density function. Let f be the density function of the normal distribution than we can write

$$\begin{aligned} S_{\vartheta}(x) \;&=\; \frac{\partial}{\partial\vartheta}\ln f_{\vartheta}(x) = \frac{\partial}{\partial\vartheta}\ln\left(\frac{1}{\sqrt{2\pi\sigma}}\,\exp-\frac{(x-\vartheta)^2}{2\sigma}\right) \\ &=\; \frac{\partial}{\partial\vartheta}\ln\left(\frac{1}{\sqrt{2\pi\sigma}}\right)-\frac{(x-\vartheta)^2}{2\sigma} \;=\; \frac{x-\vartheta}{\sigma}. \end{aligned}$$

So the variance of S_{ϑ} and thus the Fisher-Information can be computed as

$$I(\vartheta) \;=\; Var(S_{\vartheta}) \;=\; Var\left(\frac{x-\vartheta}{\sigma}\right) \;=\; \frac{1}{\sigma^2}Var(x)+\frac{1}{\sigma^2}Var(\vartheta) \;=\; \frac{1}{\sigma^2}Var(\vartheta) \;=\; \frac{\sigma}{\sigma^2} \;=\; \frac{1}{\sigma}.$$

□

Therefore, referring to Carpenter and Kenward (2008, p. 79), by using (3.13) we can write our percentage of missing information as

$$\left(\frac{\hat{W}+\hat{B}}{\hat{W}}\right)^{-1} 100\% \;=\; \frac{\hat{W}}{\hat{W}+\hat{B}}\,100\%. \tag{3.16}$$

However, they note, that for practical work (3.16) is preferred since it is sufficient enough and intuitive. However Rubin (1987, p. 93) has given a better expression to estimate the fraction of missing information. Using the Fisher-Information, we obtain

$$I_D(\vartheta) \;=\; I_D \;=\; (\nu+1)(\nu+3)^{-1}\hat{V}^{-1}, \quad D<\infty.$$

Thus, when we use the result from above, we can define $I_\infty := \hat{W}^{-1}$ and write the fraction of missing information as

$$\gamma_D \;=\; \frac{I_\infty - I_D}{I_\infty} \;=\; \frac{r_m + 2/(\nu+3)}{r_m+1},$$

according to Rubin (1987, p. 93), with $r_m = (1 + D^{-1})\hat{B}/\hat{W}$. Similarly, the fraction of information is

$$\frac{I_D}{I_\infty} = \left(1 + \frac{\gamma_D}{D}\right)^{-1}. \tag{3.17}$$

(Nassiri et al., 2020, p. 126). Using (3.17), it can be determined how many imputations are needed to cover a certain level of information. If, for example, 10% of data is missing, $D = 3$ imputations would be enough, to get about 97% of information in comparison to I_∞.

Royston (2004, p. 236-240) on the other hand suggested a different approach. He investigated the relation between the length of the confidence interval for $\sqrt{\hat{V}_D}\, t_\nu$ (the confidence coefficient, CC) and D and furthermore, what effect D has on the coefficient of variation (CV; standard deviation divided by the mean of CC). He found that both the length of the confidence interval and the CV is falling in D. Based on this findings he formulated a rule of thumb, that D should be selected in a manner that the CV for the worst-case parameter is less than α, the given Type I error rate. In addition to that he articulated an equivalent criterion that demands, that the standard deviation of $ln(CC)$ is, like in the first criterion, less than α (Royston, 2004, p. 239).

In the next section we want to take a look at an iterative procedure for finding estimators based on an optimal number of imputations.

4 The iterative Multiple Imputation Procedure

The algorithm we are about to examine was developed by Nassiri et al. (2020). Like we have seen in the previous section so far characteristics given a certain D have been compared to characteristics for $D \longrightarrow \infty$, to determine an optimal number of imputations. Nassiri et al. (2020) on the other hand propose a slightly different approach. They treat it as an convergence problem, where $\bar{\vartheta}_D$ is compared to $\bar{\vartheta}_{D+1}$, since the estimated parameter and its variance will converge at some point (Nassiri et al., 2020, p. 126). But at first we want to briefly review some distances between vectors which are potential candidates to be used in the iterative multiple imputation procedure (imi).

4.1 Preliminaries

To get an idea if the estimator converges, we need to know how much it differs from its successor. Since our estimator can be assumed to be a vector in general, it is obvious to use norms to measure the distance between two vectors. The best known family of norms are the so-called l_p-norms of a vector $x \in \mathbb{R}^n$ which are defined as

$$\|x\|_p := \left(\sum_{k=1}^{n} |x_k|^p\right)^{1/p}. \tag{4.1}$$

Notable members of this family are the Euclidean norm or l_2-norm

$$\|x\|_2 := \left(\sum_{k=1}^{n} |x_k|^2\right)^{1/2} = \sqrt{x^T x},$$

and the maximum norm

$$\|x\|_\infty := max\{|x_1|, ..., |x_n|\}.$$

With the above norms we can define the distance between $\bar{\vartheta}_D$ and $\bar{\vartheta}_{D+1}$ as

$$\begin{aligned} d_{D+1}^{Euc} &= \sqrt{(\bar{\vartheta}_{D+1} - \bar{\vartheta}_D)^T \, (\bar{\vartheta}_{D+1} - \bar{\vartheta}_D)} \\ d_{D+1}^{\infty} &= max\{|\bar{\vartheta}_{D+1} - \bar{\vartheta}_D|\}. \end{aligned}$$

Another potential candidate is the so-called Mahalanobis distance (MD) (Mahalanobis, 1936), which is similar to the Euclidean distance (ED) but with the difference that the inverse of a covariance matrix, S, is inserted between $\bar{\vartheta}_D$ and its successor:

$$d_{D+1}^{Mah} = \sqrt{(\bar{\vartheta}_{D+1} - \bar{\vartheta}_D)^T \, S^{-1} \, (\bar{\vartheta}_{D+1} - \bar{\vartheta}_D)}. \tag{4.2}$$

The intuition of multiplying with the inverse of the covariance matrix is to correct for the correlation within the data (De Maesschalck et al., 2000, p. 5). Consider the case of a 2-dimensional vector $x = (x_1, x_2)^T$ with the objective to determine the Mahalanobis distance to $\bar{x} = (\bar{x_1}, \bar{x_2})$ and with the variance $\sigma = (\sigma_1, \sigma_2)$. De Maesschalck et al. (2000, p. 5) show, that the MD for this vector can be written as

$$d_{D+1}^{Mah} = \sqrt{\left(\frac{x_1 - \bar{x_1}}{\sigma_1}\right)^2 + \left[\left(\left(\frac{x_2 - \bar{x_2}}{\sigma_2}\right) - \varrho_{12}\left(\frac{x_1 - \bar{x_1}}{\sigma_1}\right)\right)\frac{1}{\sqrt{1 - \varrho_{12}^2}}\right]^2}$$

with ϱ_{12} being the correlation between the first and second component. It can be seen, that for $\varrho_{12} = 0$ we basically get the Euclidean distance for normalized vectors. Graphically, this would mean that a non-zero correlation produces an ellipse. If we have a positive correlation and two random vectors with the same MD to $\bar{x}$, the vector that lies in the less correlated direction, let us say y, would have a smaller MD than the other, z. According to De Maesschalck et al. (2000, p. 5) this can be interpreted that the probability, that an additional vector lies in the surrounding of y is less than for lying near z. Thus, the MD assigns y a larger distance (De Maesschalck et al., 2000, p. 5).

In our case of multiple imputation, $\hat{W}_{D+1}$ and $\hat{V}_{D+1}$ can be potentially used as S among others.

4.2 The Algorithm

Nassiri et al. (2020, p. 126) formulate the imi algorithm as follows:

1. **Start.** Select an initial number of imputed datasets, D_0, $\bar{\vartheta}_{D_0} = \sum_{i=1}^{D_0} \hat{\vartheta}_i / D_0$
2. **Update.** For $D > D_0$,

$$\bar{\vartheta}_{D+1} = \frac{D \, \bar{\vartheta}_D + \hat{\vartheta}_{D+1}}{D+1} \tag{4.3}$$

3. **Distance.** Compute: $d_{D+1} = d(\bar{\vartheta}_{D+1}, \bar{\vartheta}_D)$ using an appropriate distance.
4. **Stopping rule.** $d_j < \varepsilon$ for $j = D+1, ..., D+k_0$

D_0 denotes the minimum number of imputations that has to be chosen in advanced. This comes in handy if it is known beforehand, that it will take a minimum number of imputations (Nassiri et al., 2020, p. 126). If this is not the case, a small value of $D_0 = 2$ leads to the evaluation of the stopping rule from the very beginning. For the stopping rule, two more parameters have to be chosen beforehand. The first is $\varepsilon > 0$ and the second $k_0 \in \mathbb{N}$. The k_0 denotes the number of iterations and the distance between the estimator $\bar{\vartheta}_{D+1}$ and its predecessor has to be smaller than ε, before the algorithm terminates. The necessity of k_0 is due to the fact, that the convergence does not have to be monotone. So it may well be that even-though in one iteration the distance satisfies the requirement of the stopping rule, the distance in the following few iterations is again well above ε. To avoid those local minima, a k_0-fold validation is quite reasonable.

The convergence rate depends on the sample size, the dimension of ϑ and the third moment of the elements of ϑ according to Gotze (1991) and Nassiri et al. (2020, p. 127). Obviously, a larger D favor the convergence while a higher percentage of missing data does not. In addition to that a higher dimension of the parameter vector make the convergence slower, since there are more parameters to estimate. Moreover, does the model have an influence on the convergence rate (Nassiri et al., 2020, p. 127). Since the imi algorithm is an iterative procedure, all those aspects are automatically considered.

The last feature that has to be determined in advance, is the choice of distance. All distances that have been mentioned in section (4.1) come into question for this. If one decide for the Mahalanobis distance (4.2), in addition to that a S must be set.

Consider

$$\bar{\vartheta}_{D+1} - \bar{\vartheta}_D \overset{(4.3)}{=} \frac{D\,\bar{\vartheta}_D + \hat{\vartheta}_{D+1}}{D+1} - \frac{(D+1)\,\bar{\vartheta}_D}{D+1} = \frac{\hat{\vartheta}_{D+1} - \bar{\vartheta}_D}{D+1}. \tag{4.4}$$

Since $\hat{\vartheta}_{D+1}$ and $\bar{\vartheta}_D$ are independent, we get can write the covariance of the difference of $\bar{\vartheta}_{D+1}$ and $\bar{\vartheta}_D$ as

$$\begin{aligned} Cov\left(\bar{\vartheta}_{D+1} - \bar{\vartheta}_D\right) \overset{(4.4)}{=} Cov\left(\frac{\hat{\vartheta}_{D+1} - \bar{\vartheta}_D}{D+1}\right) &= \frac{1}{(D+1)^2}[Var(\hat{\vartheta}_{D+1}) + Var(\bar{\vartheta}_D)] \\ &\overset{(??)}{=} \frac{1}{(D+1)^2}[Var(\hat{\vartheta}_{D+1}) + \hat{W}_D)]. \end{aligned} \tag{4.5}$$

Because of the division by $D + 1$, Nassiri et al. (2020, p. 127) consider this not to be suitable, as it changes with each iteration. In contrast to that, a good choice for a distance should be '*sensitive to the fraction of missing data, but robust against rescaling model parameters*', Nassiri et al. (2020, p. 127). In addition, their results show that $\hat{B}$ fails the sensitivity requirement, while $\hat{W}$ behaves chaotic in some cases. Thus, it is suggested to use $\hat{V}$, which provides a proper balance of the within and between variability.

5 Simulation Study

In this section we want to evaluate the algorithm we introduced in section 4.2. For this purpose, we generate normally distributed data sets consisting of random vectors $\mathbf{Y}_i$ that are distributed as following

$$\mathbf{Y}_i \sim N(\mu \mathbf{1}_n, \Sigma), \quad \text{with} \tag{5.1}$$

$$\Sigma = \sigma^2 I_n + \tau J_n, \quad \sigma^2, \tau > 0 \quad \text{and} \quad \tau = \frac{\varrho\sigma^2}{1-\varrho}. \tag{5.2}$$

In this case, $\mathbf{1}_n$ denotes a vector of length n consisting of ones, while I_n is a n-dimensional identity matrix and J_n a $n \times n$ matrix of ones. To obtain a data set we draw N random vectors which results in a $N \times n$ matrix, where the columns have a mean of $\mu = 0$ and a covariance like (5.2) for $N \longrightarrow \infty$. To simulate missing values, we use the function *ampute* from the R library *mice*. The function creates missing data under a missing at random mechanism, which is in line with the assumptions we made in the beginning of Section 3.2. The imputations on the other hand are made with another R library called *Amelia*, which is specialised in creating multiple complete data sets with a multivariate normal predictive model (Nassiri et al., 2020, p. 129).

We create scenarios by using different settings for the involved parameters. Parameter setups regarding the distribution of the random vectors are created from

- Proportion of missing data: $\{0.1, 0.7\}$
- $\varrho \in \{0.1, 0.9\}$,

while we choose algorithm specific variables from

- $\varepsilon \in \{0.005, 0.05\}$
- $k_0 \in \{1, 3, 5\}$.

Settings that do not change across the simulations are $\sigma^2 = 0.25$, the initial number of imputation will be $D_0 = 2$ and $\mu = 0$. The distance that is used is the Mahalanobis distance with $S = \hat{V}$. All scenarios are repeated 100 times.

The objective is to determine for each scenario the average number of imputations ,the imi algorithm uses until it reaches convergence, and its standard deviation. In addition to that we, observe how close the resulting estimated coefficients are to the true ones using mean absolute deviation (MAD).

In the next two subsections we briefly present the models we are going to use.

5.1 Compound-Symmetry

The first model is called compound-symmetry (CS). It is described in detail by Hermans et al. (2019). The model focuses on samples that consists of clusters of different sizes. In the context of this study the mentioned clusters are represented by the n-dimensional random vectors that have been introduced in the beginning of this chapter. In the paper Hermans et al. they derive a maximum likelihood estimator for the coefficients μ_n, σ_n^2 and τ_n. Since their paper considered vectors of different sizes, but this simulation study only uses vectors with a constant length of $n = 5$, the multiplicity of the class of n-dimensional vectors is N.

Therefore the index of the coefficients that are to be estimated can be dropped and their estimators (Hermans et al., 2019, p. 1113) can be rewritten as

$$\begin{aligned}\hat{\mu} &= \frac{1}{Nn}\sum_{i=1}^{N}\sum_{j=1}^{n} Y_{ij},\\ \hat{\sigma}^2 &= \frac{1}{Nn(n-1)}\left(n\sum_{i=1}^{N} Y_i^T Y_i - \sum_{i=1}^{N} Y_i^T J_n Y_i\right),\\ \hat{\tau} &= \frac{1}{Nn(n-1)}\left(\sum_{i=1}^{N} Y_i^T J_n Y_i - \sum_{i=1}^{N} Y_i^T Y_i\right).\end{aligned}$$

According to Molenberghs et al. (2011, p. 9f) the variance of $\hat{\mu}$ can be written as

$$var(\hat{\mu}) = \frac{\sigma^2 + n\tau}{Nn}, \tag{5.3}$$

while the covariance matrix of $\hat{\sigma}^2$ and $\hat{\tau}$ is given by

$$\begin{pmatrix}\hat{\sigma}^2\\ \hat{\tau}\end{pmatrix} = \frac{2\sigma^4}{Nn(n-1)}\begin{pmatrix} n & -1\\ -1 & \frac{\sigma^4+2(n-1)\tau\sigma^2+n(n-1)\tau^2}{\sigma^4}\end{pmatrix} \tag{5.4}$$

Since μ is independent of the other two variance related coefficients (Hermans et al., 2019, p. 1113), $\hat{W}$, the within variability that is defined in section 3.2.1, is obtained by combining (5.3) and (5.4) to one 3×3 covariance matrix.

5.2 Logistic Regression

The logistic regression is a method, which is usually used for classification. It maps an input vector to a value in $[0, 1]$ which can be interpreted, depending on the context, as probability $Pr(Y = 1|X\beta)$. In contrast to the linear regression function, the logistic regression function is not linear but rather s-shaped with the already mentioned convergence against 1 and 0 for $+\infty$ and $-\infty$, respectively. The logistic regression function is given by James et al. (2013, p. 134):

$$p(X) = \frac{\exp(X\beta)}{1+\exp(X\beta)} = \frac{1}{1+\exp(-X\beta)}, \tag{5.5}$$

where X is the $N \times p$ design matrix and β the p-dimensional coefficient vector, with p denoting the number of predictors, which in this case is three. To create the data set, we again generate N random vectors according to (5.1) but this time with size 2. We obtain a 100×2 matrix which represents the design matrix X after we add a one-column for the constant. Furthermore, we choose $\beta = (0.2, -2, 0.5)^T$ and create the response variables by inserting X and the selected β into (5.5). As mentioned in the introduction of this section, *ampute* and *Amelia* is used to get the imputed data sets. Maximum likelihood is used to estimate the coefficients $\hat{\beta}$. To fit the logistic regression function, the *glm* R function is used.

In addition to that, the variances and covariances can also be obtained with the maximum likelihood estimations. Consider a matrix with the negative of the second partial derivatives of the log-likelihood function:

$$
\begin{aligned}
\frac{\partial^2 L(\beta)}{\partial \beta_j^2} &= -\sum_{i=1}^{N} x_{ij}^2 p_i (1-p_i) \\
\frac{\partial^2 L(\beta)}{\partial \beta_j \beta_l} &= -\sum_{i=1}^{N} x_{ij} x_{il} p_i (1-p_i)
\end{aligned}
$$

for $j, l \in \{1,...,p\}$ with $p_i = p(x_i)$ (Hosmer, 2013, p. 37). This matrix is called observed information matrix and its inverse provides the asymptotic covariance matrix for β. However, for the simulation study we use the R function *vcov()* from the package *stats* to calculate the covariance matrix for the parameters.

5.3 Simulation Results

The objective of this simulation study is to supplement the simulation of Nassiri et al. (2020) with some more parameter settings and a metric for how close the estimators actually get to the true parameters using the algorithm. Latter has not yet been studied.

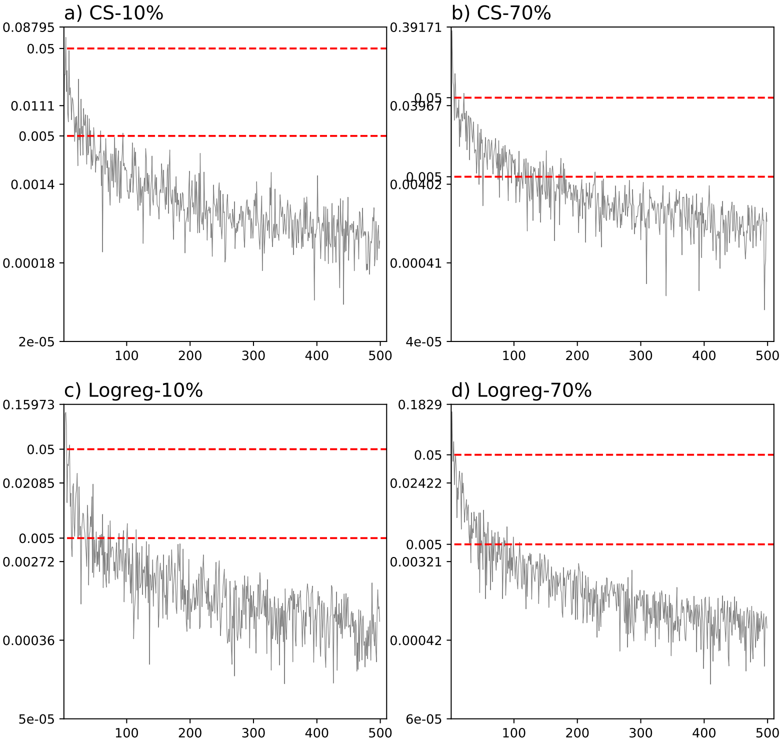

Figure 1: Convergence rates of the distances for CS and Logreg for $\sigma^2 = 0.25$, $\varrho = 0.1$, $k_0 = 5$ and $\beta = (0.2, -2, 0.5)^T$.

When we take a look at the tables 1-3, we can see that for a moderate proportion of missing data and $\varepsilon = 0.05$ the average number of imputations, the algorithm determines, is slightly above five, which is still in line with suggestions from literature (see Rubin, 1987) or the

			$k_0 = 1$				
Model	ϱ	ε	Mean	SD	$\mu(\beta_1)$ MAD	$\sigma^2(\beta_2)$ MAD	$\tau(\beta_3)$ MAD
CS-10%	0.1	0.005	25.55	10.11	0.02	0.02	0.01
		0.05	4.98	1.66	0.02	0.01	0.01
	0.9	0.005	12.86	4.79	0.14	0.02	0.29
		0.05	3.7	1.01	0.12	0.01	0.26
CS-70%	0.1	0.005	53.58	16.97	0.03	0.02	0.01
		0.05	10.28	3.52	0.03	0.02	0.01
	0.9	0.005	25.42	9.33	0.13	0.01	0.26
		0.05	5.72	2.49	0.12	0.02	0.25
Logreg-10%	0.1	0.005	22.57	9.01	0.11	0.14	0.06
		0.05	4.82	1.6	0.10	0.13	0.05
	0.9	0.005	22.16	9.26	0.11	0.15	0.11
		0.05	4.93	1.79	0.1	0.16	0.11
Logreg-70%	0.1	0.005	22.21	8.57	0.1	0.14	0.06
		0.05	4.91	1.46	0.1	0.15	0.05
	0.9	0.005	24.45	9.92	0.1	0.17	0.11
		0.05	4.65	1.48	0.1	0.15	0.11

Table 1: Mean, SD and their mean absolute deviation from the true parameter (β_i corresponds to the Logreg model and μ, σ^2, τ the CS model) for selected D given $\sigma^2 = 0.25$, $k_0 = 1$ and different values for ε and ϱ using the Mahalanobis distance with $S = \hat{V}$.

calculation example for (3.17) for CS as well as Logreg. The selection of k_0 does not seem to influence outcome heavily.

In contrast to that the choice of k_0 heavily influences the mean and SD in the case of $\varepsilon = 0.005$. This makes sense when we take a look at Figure 1. The distances converge really fast at the beginning and drop more or less immediately below the 0.05 mark and it gets quite flat right at the 0.005 level, especially for CS-70%. This is reflected in the highest standard deviation among those four cases displayed on Figure 1. This is similar to the maximization of functions where functions with a maximum located in a steep area are preferred, since the maximum can be located more precisely.

Intuitively, the MAD of the coefficients in the $\varepsilon = 0.005$ case should be lower than in the 0.05 case, since it generally takes more imputation for the algorithm to converge. However, in this simulation there is barely a difference. In most of the cases, the estimators with the higher choice of ε are roughly as close to the true coefficients as the estimators with the lower ε. To our surprise is the MAD for the choice of $\varepsilon = 0.05$ even lower, more than

			$k_0 = 3$				
Model	ϱ	ε	Mean	SD	$\mu(\beta_1)$ MAD	$\sigma^2(\beta_2)$ MAD	$\tau(\beta_3)$ MAD
CS-10%	0.1	0.005	52.83	14.22	0.02	0.01	0.01
		0.05	7.23	2.3	0.02	0.01	0.01
	0.9	0.005	30.06	11.11	0.12	0.01	0.31
		0.05	4.54	1.74	0.12	0.01	0.27
CS-70%	0.1	0.005	121.36	29.1	0.02	0.02	0.01
		0.05	15.99	3.69	0.02	0.02	0.01
	0.9	0.005	66.6	23.46	0.12	0.02	0.25
		0.05	9.86	3.63	0.12	0.01	0.24
Logreg-10%	0.1	0.005	46.11	17.62	0.1	0.13	0.05
		0.05	6.43	2.33	0.1	0.13	0.06
	0.9	0.005	47.35	19.28	0.11	0.15	0.11
		0.05	6.47	2.55	0.1	0.16	0.11
Logreg-70%	0.1	0.005	49.75	18.94	0.1	0.15	0.05
		0.05	6.23	2.06	0.1	0.14	0.06
	0.9	0.005	47.56	16.62	0.11	0.16	0.11
		0.05	6.17	2.59	0.11	0.16	0.11

Table 2: Mean, SD and their mean absolute deviation from the true parameter (β_i corresponds to the Logreg model and μ, σ^2, τ the CS model) for selected D given $\sigma^2 = 0.25$, $k_0 = 3$ and different values for ε and ϱ using the Mahalanobis distance with $S = \hat{V}$.

			$k_0 = 5$				
Model	ϱ	ε	Mean	SD	$\mu(\beta_1)$ MAD	$\sigma^2(\beta_2)$ MAD	$\tau(\beta_3)$ MAD
CS-10%	0.1	0.005	65.36	19.83	0.02	0.01	0.01
		0.05	7.73	2.67	0.02	0.01	0.01
	0.9	0.005	42.42	17.87	0.12	0.01	0.29
		0.05	5.36	2.43	0.12	0.01	0.23
CS-70%	0.1	0.005	152.23	27.11	0.02	0.01	0.01
		0.05	19.06	4.47	0.02	0.01	0.01
	0.9	0.005	91.55	26.09	0.11	0.02	0.24
		0.05	12.94	4.85	0.12	0.01	0.28
Logreg-10%	0.1	0.005	57.61	18.04	0.1	0.14	0.06
		0.05	6.52	2.42	0.1	0.12	0.06
	0.9	0.005	57.88	20.02	0.11	0.14	0.1
		0.05	7.56	3.43	0.11	0.16	0.12
Logreg-70%	0.1	0.005	60.93	17.99	0.1	0.14	0.05
		0.05	7.04	2.94	0.11	0.14	0.06
	0.9	0.005	60.7	22.02	0.11	0.16	0.11
		0.05	7.49	2.85	0.11	0.15	0.1

Table 3: Mean, SD and their mean absolute deviation from the true parameter (β_i corresponds to the Logreg model and μ, σ^2, τ the CS model) for selected D given $\sigma^2 = 0.25$, $k_0 = 5$ and different values for ε and ϱ using the Mahalanobis distance with $S = \hat{V}$.

once. Altogether, in our simulation study, a higher precision in terms of a lower choice of ε does not add any value and merely increases D. Therefore, at least for our simulation set up and the chosen models, the choice of $\varepsilon = 0.05$ is preferable.

Upon examining the tables, a noteworthy observation is the behavior exhibited when selecting a value of $\varrho = 0.9$, which pertains to the correlation between two random vector entries or five entries, depending on the choice of the model. When you compare the MADs of the CS-70% estimators with a correlation of 0.1 with the CS-10% estimators with a correlation of 0.9, it can be seen, that for compound symmetry rather the correlation is crucial for the MAD than the proportion of missing data. Such a relationship was not clearly evident in the logistic regression. Furthermore, the correlation seems to reduce mean of the needed number of imputation and its standard deviation.

In summary, the imi procedure produces viable estimates with a reasonable number of imputation for the correct choice of ε where 'the lower the more precise' does not necessarily hold.

6 Conclusion

In this seminar paper we studied multiple imputation, a technique to handle missing data, which is particularly useful to investigate the sensitivity of estimators. We derived the underlying theory under the assumption of ignorable mechanisms including expressions for the estimators variance depending on the number of imputations and its distribution for $D \longrightarrow \infty$ and small D. Based on this results methods have been developed to determine what the optimal number of imputations is.

From all of this methods we picked the 'iterative multiple imputation procedure' developed by Nassiri et al. (2020) for a more detailed examination. In contrast to other existing methods, imi suggests, an iterative algorithm that imputes data sets until convergence. Like most algorithms it depends on hyper-parameters for the stopping rule, the allowed distance between the parameters of two steps ε, the number of validation steps k_0 and the used distance between two consecutive estimations.

In order to test the algorithm we did a simulation study similar to Nassiri et al. (2020) but with a different choice of parameters which included a more correlated data set. In addition to that we evaluated how close the from the algorithm created estimators are to the true coefficients and what it is influenced by.

We found that the main factor behind the number of imputation is the proportion of missing data, the number of validation steps and ε, which is in line with the literature. In addition to that our study confirms the rule of thumbs that between 2 and 10 imputations are sufficient to obtain estimators with small MAD.

References

[1] Carpenter, J. R., and Kenward, M. G. (2008). *Missing Data in Clinical Trials: Practical Guide*, Birmingham: National Institute for Health Research, Publication RM03/JH17/MK.

[2] De Maesschalck, R., Jouan-Rimbaud, D., Massart, D. L. (2000). *The Mahalanobis distance*. Chemometrics and Intelligent Laboratory Systems 50, 1–18.

[3] Fisher, R. A. (1971). *The Design of Experiments* (9th ed.), Hafner Press, New York.

[4] Gotze, F. (1991). *On the Rate of Convergence in the Multivariate CLT*. The Annals of Probability 19, 724-739.

[5] Hermans, L., Nassiri, V., Molenberghs, G., Kenward, M. G., Van der Elst, W., Aerts, M., Verbeke, G. (2019). *Clusters with unequal Size: Maximum Likelihood versus weighted Estimation in large Samples*. Statistica Sinica (forthcoming).

[6] Hosmer, D. W. (2013). *Applied Logistic Regression* (3rd ed.), John Wiley & Sons, New York.

[7] James, G., Witten, D., Hastie, T., Tibshirani, R. (2013). *An introduction to statistical learning* (2nd ed.), Springer, New York.

[8] Kamps, U., Cramer, E. (2021). *Grundzüge der Stochastik* (3. Auflage), RWTH Aachen, Aachen.

[9] Mahalanobis, P. C. (1936). *On the generalized Distance in Statistics*. The journal of the Asiatic Society of Bengal 26, 541-588.

[10] Molenberghs, G., Verbeke, G., Iddi, S. (2011). *Pseudo-likelihood methodology for partitioned large and complex samples*. Statistics & Probability Letters 81, 892-901.

[11] Nassiri, V., Molenberghs, G., Verbeke, G., Barbosa-Breda, J. (2020). *Iterative Multiple Imputation: A Framework to Determine the Number of Imputed Datasets*. The American Statistician 74, 125-136.

[12] Pinto, L. A., Willekens, K., Van Keer, K., Shibesh, A., Vandewalle, E., Molenberghs, G., and Stalmans, I. (2015). *Leuven Eye Study - Baseline and Methods*. Acta Ophthalmologica 93.

[13] Robert, C. P. (2007). *The Bayesian choice: from decision-theoretic foundations to computational implementation* (Vol. 2), Springer, New York.

[14] Royston, P. (2005). *Multiple imputation of missing values*. The Stata Journal 4, 227-241

[15] Rubin, D. B. (1978). *Multiple Imputations in Sample Surveys—A Phenomenological Bayesian Approach to Nonresponse*. Proceedings of the Survey Research Methods Section of the American Statistical Association 1, American Statistical Association, 20–34

[16] Rubin, D. B. (1987). *Multiple Imputation for Nonresponse in Surveys*, John Wiley & Sons, New York.

[17] Rubin, D. B., Little, R. (2002). *Statistical Analysis with Missing Data* (2nd ed.), John Wiley & Sons, New York.

[18] Rubin, D. B., Schenker, N. (1986). *Multiple Imputation for Interval Estimation From Simple Random Samples With Ignorable Nonresponse.* Journal of the American Statistical Association 81, 366–374.