Technische Universität Wien
Vienna University of Technology

Institut f. Statistik u. Wahrscheinlichkeitstheorie
1040 Wien, Wiedner Hauptstr. 8-10/107
AUSTRIA
http://www.statistik.tuwien.ac.at

EM-based stepwise regression imputation using standard and robust methods

M. Templ, A. Kowarik, and P. Filzmoser

Forschungsbericht CS-2010-3
Oktober 2010

Kontakt: P.Filzmoser@tuwien.ac.at

EM-based stepwise regression imputation using standard and robust methods

M. Templ*, A. Kowarik†, P. Filzmoser‡

October 13, 2010

Imputation of missing values is one of the major tasks for data pre-processing in many areas. Whenever imputation of data from official statistics comes into mind, several (additional) challenges almost always arise, like large data sets, data sets consisting of a mixture of different variable types, or data outliers. The aim of this contribution is to propose an automatic algorithm called `IRMI` for iterative model-based imputation using robust methods, encountering for the mentioned challenges, and to provide a software tool in R. This algorithm is compared to the algorithm `IVEWARE`, which is the "recommended software" for imputations in international and national statistical institutions. Using artificial data and real data sets from official statistics and other fields, the advantages of `IRMI` over `IVEWARE` – especially with respect to robustness – are demonstrated.

Keywords: EM-based regression imputation, robustness, R

1 Introduction

The imputation of missing values is especially important in official statistics, because virtually all data sets from this area deal with the problem of missing information due to non-responses, or because erroneous values have been set to missing. This has especially consequences for statistical methods using the multivariate data information. The naive approach, namely omitting all observations that include at least one missing cell, is not attractive because a lot of valuable information might still be contained in these observations. On the other hand, the estimation of the missing cells can indroduce

*Dept. of Statistics and Probability Theory, Vienna University of Technology, Wiedner Hauptstr. 8-10, 1040 Vienna, Austria
†Methods Unit, Statistics Austria, Guglgasse 13, 1110 Vienna, Austria
‡Dept. of Statistics and Probability Theory, Vienna University of Technology, Wiedner Hauptstr. 8-10, 1040 Vienna, Austria

additional bias, and valid estimates and inferences can only be made if the missing data are *missing completely at random* (MCAR) [see, e.g., Little and Rubin, 1987]. Even in this case there are further challenges, and these are very typical in data sets from official statistics:

Mixed type of variables in the data: Data from official statistics typically consist of variables that have different distributions, i.e. various variables are binary scaled, some variables might be categorical, and some variables could be determined to be of continuous scale. If missing values are present in all these variable types, the challenge is to estimate the missing values based on the whole multivariate information.

Semi-continuous variables: Another challenge is the presence of variables in the data set which consist of a part with continuous scale, but also include a certain proportion of equal values (typically zeros). The distribution of such variables is often referred to as "semi-continuous" distribution [see, e.g., Schafer and Olson, 1999]. Data consisting of semi-continuous variables are, for example, income components in the *European Union Statistics of Income and Living Condition (EU-SILC)* survey, or tax components in tax data, in which one part of such a variable origins from a continuous distribution, and the other part consists of (structural) zeros.

Large data sets: Since data collection is a requirement in many fields nowadays, the resulting data sets can become "large", and thus the the computation time of imputation methods is an important issue. One might argue that many such data sets can be decomposed into subsets referring to sub-populations, which are for instance defined by the NACE-codes in *Structural Business Survey* (SBS) data. Still, these subsets can contain more that 50000 observations, which calls for fast methods for data imputation.

Far from normality: A common assumption used for multivariate imputation methods is usually that the data originate from a multivariate normal distribution, or that they can be transformed to approximate multivariate normal distribution. This is violated in presence of outlying observations in the data. In this case, standard methods can result in very biased estimates for the missing values. It is then more advisable to use robust methods, being less influenced by outlying observations [see, e.g., Beguin and Hulliger, 2008, Serneels and Verdonck, 2008, Hron et al., 2008].

Note that prior exclusion of outliers before imputation is not straightforward. For example, when regression imputation is applied, leverage points might only be detected when analyzing the residuals from robust regression but might not be reliably identified from a least-squares fit nor by other multivariate outlier detection methods.

Sampling weights: Sampling weights are typically used in official statistics. The inclusion of sampling weights in the imputation process is not relevant or should even

be avoided. It is reasonable to find good estimates for the missing values without weighting the observations by the design weights. For example, when nearest neighbor imputation is applied by incooperating the design weights, the nearest neighbors are usually those having similar weights. However, the aim is to find those observations as donors that are similar with respect to their multivariate information, independent from their actual weights. In case of regression imputation, the large weights could introduce (additional) leverage points that could result in poor imputation quality.

1.1 Imputation methods

Many different methods for imputation have been developed over the last few decades. The techniques for imputation may be divided into univariate methods such as column-wise (conditional) mean imputation, and multivariate imputation. In the latter case there are basically three approaches: distance-based imputation methods such as k-nearest neighbor imputation, covariance-based methods such as the approaches by Verboven et al. [2007] or Serneels and Verdonck [2008], and model-based methods such as regression imputation.

If an imputation method is able to deal with the randomness inherent in the data, it can be used for multiple imputation, generating more than one candidate for a missing cell [Rubin, 1987]. Multiple imputation is one way to reflect the sampling variability, but it should only be used with careful consideration of the underlying distributional assumptions and underlying models [see also Fay, 1996, Durrant, 2005]. In addition, if assumptions for the distribution of the occurrence of nonresponse are made but violated, poor results might be obtained [see also Schafer and Olsen, 1998]. The sampling variability can also be reflected by adding a certain noise to the imputed values, and valuable inference can also be obtained by applying bootstrap methods [Little and Rubin, 1987, Alfons et al., 2009]. However, most of the existing methods assume that the data originate from a multivariate normal distribution (e.g. the MCMC methods of the imputation software MICE [van Buuren and Oudshoorn, 2005], Amelia [Honaker et al., 2009], mi [Yu-Sung et al., 2009] or mitools [Lumley, 2010]. This assumption becomes inappropriate as soon as there are outliers in the data, or in case of skewed or multimodal distributions. Since this is a very frequent situation with practical data sets, imputation methods based on robust estimates are gaining increasing importance.

The basic procedure behind most model-based imputation methods is the EM-algorithm [Dempster et al., 1977], which can be thought of a guidance for the iterative application of estimation, adaption and re-estimation. For the estimation, usually regression methods are applied in a iterative manner, which is known under the names regression switching, chain equations, sequential regressions, or variable-by-variable Gibbs sampling [see, e.g., van Buuren and Oudshoorn, 2005, Muennich and Rässler, 2004].

1.2 Software for imputation

The R package **mix** by Schafer [2009, 1996] considers most of the challanges described above, but it cannot handle semi-continuous variables, although Schafer and Olson [1999] described the problems with semi-continuous variables. In addition also the R package **mice** by van Buuren and Oudshoorn [2005] is not designed to deal with semi-continuous variables. The R package **mi** [Yu-Sung et al., 2009] is well suited for multiple imputation in general, but it has the same limitations related to semi-continuous variables. This problem is treated in `IVEWARE`, a set of C and Fortran functions for which also SAS Macros are available [Raghunathan et al., 2001]. The algorithms in **mi** and `IVEWARE` are based on iterative regression imputation. **mi** starts the algorithm by a rough initialization (randomly chosen values). The same concept is used by **MICE** (Multiple Imputation by Chain Equations) van Buuren and Oudshoorn [2005] and the Amelia package of Honaker et al. [2009], where first bootstrap samples with the same dimensions as the original data are drawn, and used for EM-based imputation.

All these algorithms and procedures cannot adequately cope with data including outliers. The aim of this contribution is to develop a procedure that is competitive with the above algorithms, but has the additional feature of being robust with respect to data outliers. Since `IVEWARE` takes care of all the mentioned problems except robustness, and because this software is also recommended by EUROSTAT [see, e.g., Eurostat, 2008a], it is natural to use it as a basis for our task.[1]

A drawback of `IVEWARE` is that the exact procedure of the algorithm is not well documented. Therefore, we analyzed the software and provide a mathematical description of the algorithm in Section 2. Section 3 introduces the robust counterpart to `IVEWARE` which we call `IRMI` (Iterative Robust Model-based Imputation). Also other improvements were included in `IRMI`, like a different strategy for the initialization of the missing values. Simple comparisons of the two algorithms on two-dimensional artificial data are made in Section 4, and more detailed comparisons based on simulations are in Section 5. Applications to real data sets are provided in Section 6. Section 7 describes the software implementation of `IRMI`, and the final section concludes.

2 The algorithm `IVEWARE`

`IVEWARE` estimates the missing values by fitting a sequence of regression models and drawing values from the corresponding predictive distributions [Raghunathan et al., 2001]. Unfortunately, a detailed description of the algorithm does not exist. Raghunathan et al. [2001] provide only a rather vague outline of the functionality of this algorithm. However, this description and the analysis of the provided software tools make it possible to get a clearer picture of the functionality:

Step 1: Sort the variables according to the amount of missing values. In order to avoid complicated notation, we assume that the variables are already sorted, i.e.

[1] IVEWARE is mentioned by Eurostat in internal task force reports and presentation slides, and it is routinely used in national statistical institutions as well as in various research organizations.

$\mathcal{M}(\boldsymbol{x}_1) \leq \mathcal{M}(\boldsymbol{x}_2) \leq \ldots \leq \mathcal{M}(\boldsymbol{x}_p)$, where $\mathcal{M}(\boldsymbol{x}_j)$ denotes the number of missing cells in variable $\boldsymbol{x}_j$. Denote $I = \{1, \ldots, k\}$ as the set of indices of the variables with no missing values, and fix $r = k + 1$.

Step 2: Set $l = k + 1$.

Step 3: Denote $m_l \subset \{1, \ldots, n\}$ the indices of the observations that are **originally** missing in variable $\boldsymbol{x}_l$, and $o_l = \{1, \ldots, n\} \backslash m_l$ the indices corresponding to the observed cells of $\boldsymbol{x}_l$. Let $\boldsymbol{X}^{o_l}_{I\backslash\{l\}}$ and $\boldsymbol{X}^{m_l}_{I\backslash\{l\}}$ denote the matrices including the variables defined by the indices $I\backslash\{l\}$ (at the start $l \notin I$), and with the observations according to the observed and missing cells of $\boldsymbol{x}_l$, respectively. Additionally, the first column of $\boldsymbol{X}^{o_l}_{I\backslash\{l\}}$ and $\boldsymbol{X}^{m_l}_{I\backslash\{l\}}$ consists of ones, taking care of an intercept term in the regression problem

$$\boldsymbol{x}_l^{o_l} = \boldsymbol{X}^{o_l}_{I\backslash\{l\}}\boldsymbol{\beta} + \boldsymbol{\varepsilon} \tag{1}$$

with unknown regression coefficients $\boldsymbol{\beta}$ and an error term $\boldsymbol{\varepsilon}$.
The distribution of the response $\boldsymbol{x}_l^{o_l}$ determines the regression model used. If the response is

- *continuous*, ordinary least squares (OLS) regression is applied;
- *categorical*, polytomous or generalized logit regression is applied;
- *binary*, logistic linear regression is applied;
- *count*, a Poisson log-linear model is applied;
- *semi-continuous*, a two-stage approach is used, where in the first stage logistic regression is applied in order to decide if a constant (usually zero) is imputed or not. In the latter case the imputation is done by OLS regression based on the continuous (non-constant) part of the response.

Step 4: Estimate the regression coefficients $\boldsymbol{\beta}$ with the corresponding model from step 3, and use the estimated regression coefficients $\hat{\boldsymbol{\beta}}$ to replace the missing parts $\boldsymbol{x}_l^{m_l}$. In the continuous case this is done by

$$\hat{\boldsymbol{x}}_l^{m_l} = \boldsymbol{X}^{m_l}_{I\backslash\{l\}}\hat{\boldsymbol{\beta}} + \tilde{\boldsymbol{\epsilon}} \quad , \tag{2}$$

where $\tilde{\epsilon}$ is drawn from $N(0, \tilde{\sigma} \cdot u)$ with $\tilde{\sigma}$ being the square of the sum of squared residuals, and u is generated from a χ^2 distribution with df degrees of freedom[2].
In the discrete case, the values for $\hat{\boldsymbol{x}}_l^{m_l}$ are chosen slightly differently [for details, see the appendix of Raghunathan et al., 2001].

Step 5: Update I in Step 1 and l in Step 2.

Step 6: Carry out Steps 2–5 until all values have been imputed.

Step 7: Set $I = \{1, \ldots, p\}$ and repeat Steps 2–4 in turn for each $l = r, r + 1, \ldots, p$, for a prespecified number of rounds, or until stable imputed values occur.

[2]The exact choice of *df* could not be verified.

The "inner loop" (Steps 1–6) works as an initialization of the missing values. In the "outer loop" (Step 7, i.e. Steps 2–4), the values that were originally missing in a response variable are re-estimated using all the (completed) remaining variables as predictors. From this description and the description given in Raghunathan et al. [2001] it is obvious that this algorithm needs at least one variable to be fully observed. However, the currently available `SAS` macro allows to have missings in all variables. If only few variables are fully observed the algorithm may provide weak results, since the initialization at the start of the "inner loop" might be poor and this might affect the imputation carried out in the "outer" loop.

3 The algorithm `IRMI`

The algorithm called `IRMI` for Iterative Robust Model-based Imputation has been implemented as function `irmi()` in the R package **VIM**. Basically it mimics the functionality of `IVEWARE` [Raghunathan et al., 2001], but there are several improvements with respects to the stability of the initialized values, or the robustness of the imputed values. Moreover, the algorithm does not require at least one fully observed variable. In each step of the iteration, one variable is used as a response variable and the remaining variables serve as the regressors. Thus the "whole" multivariate information will be used for imputation in the response variable. The proposed iterative algorithm can be summarized as follows:

Step 1: Initialize the missing values using a simple imputation technique (e.g. k-nearest neighbor or mean imputation (default)).

Step 2: Sort the variables according to the original amount of missing values. We now assume that the variables are already sorted, i.e. $\mathcal{M}(\boldsymbol{x}_1) \geq \mathcal{M}(\boldsymbol{x}_2) \geq \ldots \geq \mathcal{M}(\boldsymbol{x}_p)$, where $\mathcal{M}(\boldsymbol{x}_j)$ denotes the number of missing cells in variable $\boldsymbol{x}_j$. Set $I = \{1, \ldots, p\}$.

Step 3: Set $l = 1$.

Step 4: Denote $m_l \subset \{1, \ldots, n\}$ the indices of the observations that were originally missing in variable $\boldsymbol{x}_l$, and $o_l = \{1, \ldots, n\} \backslash m_l$ the indices corresponding to the observed cells of $\boldsymbol{x}_l$. Let $\boldsymbol{X}^{o_l}_{I\backslash\{l\}}$ and $\boldsymbol{X}^{m_l}_{I\backslash\{l\}}$ denote the matrices with the variables corresponding to the observed and missing cells of $\boldsymbol{x}_l$, respectively. Additionally, the first column of $\boldsymbol{X}^{o_l}_{I\backslash\{l\}}$ and $\boldsymbol{X}^{m_l}_{I\backslash\{l\}}$ consists of ones, taking care of an intercept term in the regression problem

$$\boldsymbol{x}_l^{o_l} = \boldsymbol{X}^{o_l}_{I\backslash\{l\}}\boldsymbol{\beta} + \boldsymbol{\varepsilon} \tag{3}$$

with unknown regression coefficients $\boldsymbol{\beta}$ and an error term $\boldsymbol{\varepsilon}$.
The distribution of the response $\boldsymbol{x}_l^{o_l}$ is considered in each regression fit. If the response is

- *continuous*, a robust regression method (see below) is applied;

- *categorical*, generalized linear regression is applied (optionally, a robust method [see Cantoni and Ronchetti, 2001] can be selected);
- *binary*, logistic linear regression is applied (optionally, a robust method [see Cantoni and Ronchetti, 2001] can be selected);
- *semi-continuous*, a two-stage approach is used, where in the first stage logistic regression is applied in order to decide if a constant (usually zero) is imputed or not. In the latter case the imputation is done by robust regression based on the continuous (non-constant) part of the response.
- *count*, robust generalized linear regression of family Poisson is used Cantoni and Ronchetti [2001].

Optionally, it is possible to use a stepwise model selection by AIC (parameter `step` in function `irmi`) to include only the most important k variables, $k \subset I \backslash \{l\}$, in the regression problem in Equation (3). Otherwise, $k = I \backslash \{l\}$.

Step 5: Estimate the regression coefficients $\boldsymbol{\beta}$ with the corresponding model in Step 4, and use the estimated regression coefficients $\hat{\boldsymbol{\beta}}$ to replace the missing parts $\boldsymbol{x}_l^{m_l}$ by

$$\hat{\boldsymbol{x}}_l^{m_l} = \boldsymbol{X}_k^{m_l} \hat{\boldsymbol{\beta}}. \tag{4}$$

Step 6: Carry out Steps 4–5 in turn for each $l = 2, \ldots, p$.

Step 7: Repeat Steps 3–6 until the imputed values stabilize, i.e. until

$$\sum_i (\hat{\boldsymbol{x}}_{l,i}^{m_l} - \tilde{\boldsymbol{x}}_{l,i}^{m_l})^2 < \delta, \qquad \text{for all } i \in m_l \text{ and } l \in I,$$

for a small constant δ, where $\hat{\boldsymbol{x}}_{l,i}^{m_l}$ is the i-th imputed value of the current iteration, and $\tilde{\boldsymbol{x}}_{l,i}^{m_l}$ is the i-th imputed value from the previous iteration.

Although we have no proof of convergence, experiments with real and artificial data have shown that the algorithm usually converges in a few iterations, and that already after the second iteration no significant improvement is obtained.

Optionally, the estimation of the regression coefficients in Step 4 can be done in the classical way. In this case the algorithm is denoted by `IMI`. The abbreviation `IRMI` refers to robust regressions as outlined in Step 4, which reduce the influence of outlying observations for estimating the regression parameters [see, e.g., Maronna et al., 2006]. Within our implementation it is possible to use least trimmed squares (LTS) regression [Rousseeuw and Van Driessen, 2002], MM-estimation (default) [Yohai, 1987] and M-estimation [Huber, 1981] whenever the response is continuous or semi-continuous. If the response variable is binary, a robust generalized linear model with family *binomial* is applied [Cantoni and Ronchetti, 2001]. When the response variables is categorical, a multinomial model is chosen, which is based on neural networks [for details, see Venables and Ripley, 2002, Ripley, 1996].

Note that robust regression also protects against poorly initialized missing values, because the estimation of the regression coefficients is based only on the majority of the observations.

The function `irmi()` also provides the option to add a random error term to the imputed values, creating the possibility for multiple imputation. The error term has mean 0 and a variance corresponding to the (robust) variance of the regression residuals from the observations from the observed response. To provide adequate variances of the imputed data the error term has to be multiplied by a factor $\sqrt{1+\frac{1}{n}\#m^l}$, considering the amount of missing values ($\#m^l$) in the response (additionally, the level of noise can be controlled by a scale parameter, which is by default set to 1). Conceptionally, this is different from all other implementations of EM-based regression imputation methods. It's somehow a simplification because only expected values are used to update former missing values until convergency. However, it guarantees much faster convergency and full control of the convergency of the algorithm. Note that to keep track of convergence of sequential methods used in `IVEWARE`, `mice` or `mi` is rather difficult because in each step predictive values are used to update former missing values instead of expected values like in `IRMI`. Within `IRMI`, errors to provide correct (co-)variances are included in a final iteration in an adequate way. The chosen factor allows multiple imputation with proper coverage rates as well (see Section 3.1). Within practical application using real-world data this approach is preferable and the results shows correct variances and coverage rates, i.e. this change of paradigm has a lot of advantages when working with complex data sets from official statistics, for example.

3.1 Properties

The imputation method should be "proper", i.e. to incorporate the variability that affects the imputed value, in order to lead to consistent standard errors [see, e.g., Rubin, 1987]. While a mathematical proof whether a complex robust imputation method is proper in Rubin's sense is virtually impossible, the problem can be addressed by Monte Carlo simulation studies. Raessler and Münnich [2004] give a detailed description on how to use simulations to determine if a multiple imputation method is proper or at least approximate proper. We investigated this problem by reproducing the simulation study given in Raessler and Münnich [2004]. Let

$$(AGE, INCOME) \sim N\left(\begin{pmatrix}40\\1500\end{pmatrix}, \begin{pmatrix}10 & 44\\44 & 300\end{pmatrix}\right)$$

the universe for which samples of size 2000 are drawn, whereas variable AGE is recoded in 6 categories. We set 30% of the income values to missing values using MCAR, MAR and MNAR mechanisms[3] [see, e.g., Little and Rubin, 1987]. Then the three data sets are imputed using `IVEWARE`, `IMI` and `IRMI`, whereas 10 multiple imputed data sets are generated. The results were combined by well known rules [Rubin, 1987]. The whole procedure is repeated 2000 times and the coverage rate is counted, which is defined as the ratio between the amount of how often the true mean is covered by the estimated confidence intervals and the number of replications (2000).

[3] Under MCAR the missing values are generated completely at random, under MAR, income is missing with higher probability the higher the value of AGE, under MNAR the probability of missing in INCOME is higher the higher INCOME.

Table 1: Coverage rates using complete case analysis and different imputation methods and different missing values mechanisms.

Missing	mPop	CC	Mean	IVEWARE	IMI	IRMI-MM
none	0.96					
MCAR		0.954	0.818	0.916	0.914	0.902
MAR		0.709	0.527	0.904	0.894	0.882
MNAR		0.822	0.820	0.918	0.916	0.906

Table 1 shows the coverage rate from complete case analyse (CC), mean imputation, `IVEWARE`, `IMI` and `IRMI-MM` (using MM-regression for robust imputation of continuous and semi-continuous variables). The value of `mPop` shows the coverage rate without any values marked as missing. For all other methods missing values are generated with different kinds of missing values mechnisms.

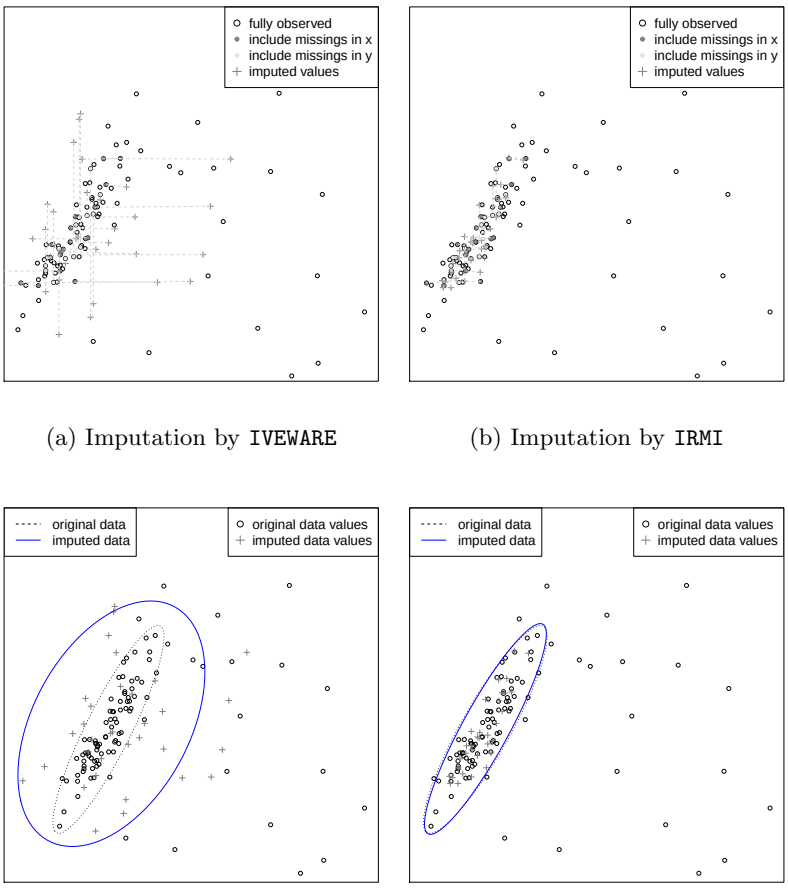
Whenever possible outliers are replaced with missing values, non-outliers will be imputed. This leads to slightly smaller confidence intervals. Therefore the coverage rate should be slightly below 0.95. All sequential imputation methods lead to comparable results and even in a MAR situation the coverage rate is reasonable. However, the coverage rate from complete analysis and mean impuation is quite low, especially in MAR situations.

4 Comparison using exploratory examples including outliers

The different functionalities of `IVEWARE` and `IRMI` can be investigated by simple data configurations where the structure is clearly visible. Here we focus on the robustness aspect of `IRMI`, and thus on the effect of outlying observations in the data.

In Figure 1, two-dimensional data consisting of continuous scaled variables are shown. A complete data set including outliers is generated, and then some values of the non-outlying part are set to be missing. The dashed lines in the upper plots (Figure 8(a) and 8(b)) join the original values with their imputed ones. It is apparent that `IVEWARE` is highly influenced by outlying observations (Figure 8(a)) while `IRMI` leads to imputed values in the central part of the point cloud (Figure 8(b)). This becomes again visible when comparing the 95% tolerance ellipses, constructed by the non-outlying (and imputed) part of the data. A 95% tolerance ellipse covers (theoretically) 95% of the observations in case of two-dimensional normal distribution. The non-robust imputation by `IVEWARE` results in an inflated tolerance ellipse when compared to the tolerance ellipse using the original complete outlier-free data (Figure 1(c)). In contrast, the robust imputation by `IRMI` causes both ellipses to be almost indistinguishable, and thus `IRMI` generates practically the same bivariate data structure (Figure 1(d)).

Figure 2 already shows both the imputation results and the 95% tolerance ellipses from another two-dimensional data set, where the variable on the vertical axis originates from a semi-continuous distribution. This situation is very typical for data from official statistics. The figures also contain boxplots in the margins of the horizontal axes, provid-



(a) Imputation by `IVEWARE`

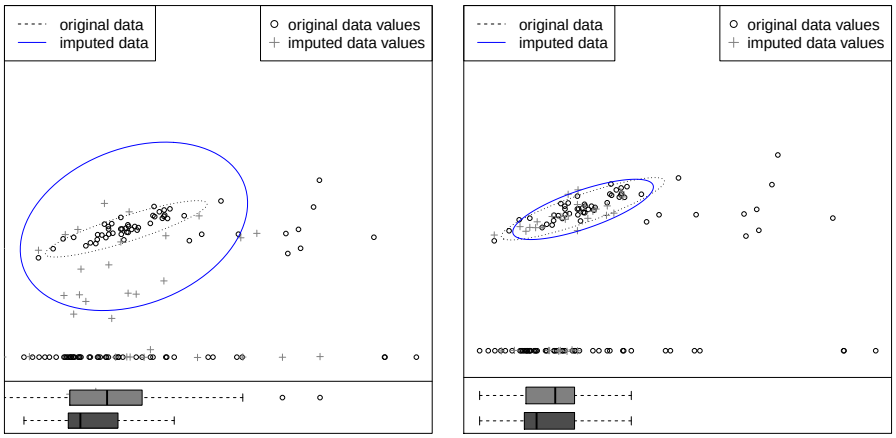
(b) Imputation by `IRMI`

(c) Imputation by `IVEWARE`

(d) Imputation by `IRMI`

Figure 1: Imputation for continuous scaled two-dimensional data by `IVEWARE` and `IRMI`. Upper plots: The dashed lines join the original values with their imputations. Lower plots: 95% tolerance ellipses characterizing the multivariate data structure of the non-outlying part of the data. The imputation by `IVEWARE` is sensitive to the outliers, while `IRMI` succeeds to impute according to the original data structure.

ing information of both the distribution of the original (dark-grey colored boxes) and the imputed (light-grey colored boxes) constant data part. The 95% tolerance ellipses are based on the continuous and non-outlying part of the data. Although `IVEWARE` should be able to cope with semi-continuous variables[4], Figure 2(a) shows that the imputed values are influenced by the outliers and the constant data part. In contrast, `IRMI` works as expected (see Figure 2(b)).



(a) Imputation by `IVEWARE` (b) Imputation by `IRMI`

Figure 2: Imputation results by `IVEWARE` and `IRMI` for a two-dimensional data set consisting of a continuous scaled and a semi-continuous variable. The covariance structure of the non-outlying continous part of the data is visualized by 95% tolerance ellipses. The constant data part is summarized by boxplots for the original (lower boxplot) and the imputed (upper boxplot) data.

Similar experiments were made with other data configurations (for example, one binary scaled variable versus one continuous scaled variable) and with other choices of the means and covariances for generating the data. The conclusions are analoguous. A more detailed comparison of `IVEWARE` and `IRMI` will be provided by simulation studies in the next section.

5 Simulation studies

For all simulations presented in this section we randomly generate data with $n = 500$ observations and p variables from a multivariate normal distribution. The population mean of (the non-outlier part of) each variable is fixed with 10. Based on the multivariate normal distribution, variables with binary and semi-continuous scale (as many other

[4]In the program `IVEWARE` we have used the data type `mixed` and the default values for all other parameters.

authors [see, e.g., Raghunathan et al., 2001] we do not consider the multinomial variables because within extensive simulation studies the computation time would then grow up a lot) are constructed by the following procedures:

Binary scale: A binary variable y with values $y_1, \ldots, y_n$ is created at the basis of a variable x with values $x_1, \ldots, x_n$ from the generated multivariate data by

$$y_i = \begin{cases} 0 & \text{with} \quad P(y_i = 0) = 1 - F_{N(\mu,\sigma^2)}(x_i) \\ 1 & \text{with} \quad P(y_i = 1) = 1 - P(y_i = 0) = F_{N(\mu,\sigma^2)}(x_i) \end{cases} \quad ,$$

for $i = 1, \ldots, n$. $F_{N(\mu,\sigma^2)}$ denotes the distribution function of x, a normal distribution with mean μ and variance σ^2. Hence, if x_i is high (low) the probability that y_i becomes zero is high (low). Depending on the choice of μ the ratio of zero and ones differs (default 50% zeros at the average)

Semi-continuous scale: Without loss of generality, we set the constant part of the variable with semi-continuous scale to zero. We use two variables from the mutlivarite data. One variable is used to generate a binary variable y with values $y_1, \ldots, y_n$. This is done in the same way as above for binary scale. A second variable $\tilde{x}$ with values $\tilde{x}_1, \ldots, \tilde{x}_n$ determines the non-constant part of the semi-continuous variable z with values $z_1, \ldots, z_n$ by

$$z_i = \begin{cases} 0 & \text{if} \quad y_i = 0 \\ \tilde{x}_i & \text{if} \quad y_i = 1 \end{cases}$$

for $i = 1, \ldots, n$.

These procedures allow that the correlation structure generated for the multivariate normally distributed data is also reflected by the variables with mixed scale.

In order to avoid complicated notation, the resulting data values are denoted by x_{ij}^{orig}, with $i = 1, \ldots, n$ and $j = 1, \ldots, p$, and the imputed values by x_{ij}^{imp}.

5.1 Error measures

The use of variables with different scale has also consequences for an error measure, providing information on the quality of the imputed data. A solution is to use different measures for categorical and binary variables, and for continuous and semi-continuous variables.

Error measure for categorical and binary variables: This error measure is defined as the proportion of imputed values taken from an incorrect category on all missing categorical or binary values:

$$err_c = \frac{1}{m_c} \sum_{j=1}^{p_c} \sum_{i=1}^{n} \mathbb{I}(x_{ij}^{orig} \neq x_{ij}^{imp}) \quad , \tag{5}$$

with $\mathbb{I}$ the indicator function, m_c the number of missing values in the p_c categorical variables, and n the number of observations.

Error measure for continuous and semi-continuous variables: The two different situations continuous and semi-continuous have to be distinguished. For the continuous parts we use the abolute relative error between the original and the imputed value. For the categorical (constant) part we count the number of incorrect categories, similar to Equation (5). Here we assume that the constant part of the semi-continuous variable is zero. Hence, the joint error measure is

$$err_s = \frac{1}{m_s} \sum_{j=1}^{p_s} \sum_{i=1}^{n} \left[\left| \frac{(x_{ij}^{orig} - x_{ij}^{imp})}{x_{ij}^{orig}} \right| \cdot \mathbb{I}(x_{ij}^{orig} \neq 0 \ \wedge \ x_{ij}^{imp} \neq 0) + \mathbb{I}((x_{ij}^{orig} = 0 \ \wedge \ x_{ij}^{imp} \neq 0) \ \vee \ (x_{ij}^{orig} \neq 0 \ \wedge \ x_{ij}^{imp} = 0)) \right] \quad (6)$$

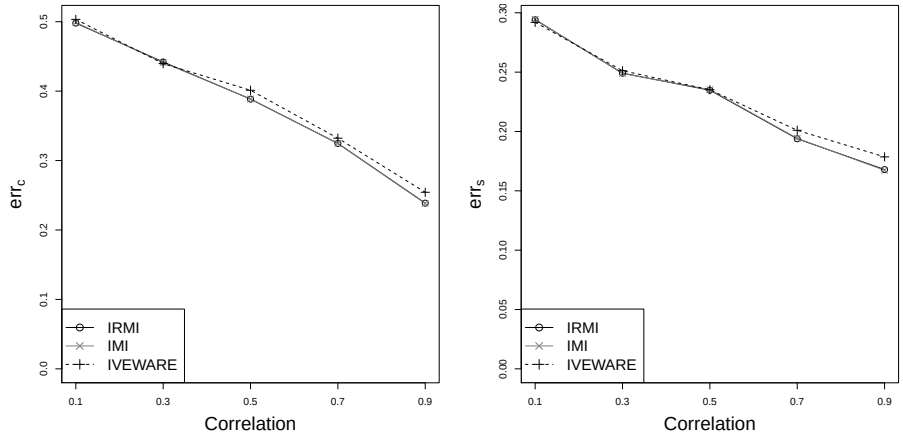
with m_s the number of missing values in the p_s continuous and semi-continuous variables. For continuous variables we assume that both the original value and the imputed value are different from zero, and thus the first part of Equation (5) measures the imputation error. In the other case, if either the original or the imputed value is zero, the second part of the equation is used.

To consider other error measures with respect to the (co-variances) of the data and errors from model predictions are out of scope within this paper. Further investigations using other error measures might be made in future.

5.2 First configuration: varying the correlation structure

In a first simulation setting we want to study the effect of the correlation structure between the variables. Therefore, the covariance matrix of the underlying multivariate normally distributed data is taken as a matrix with variances of one in the main diagonal, and otherwise constant values. These are chosen in 4 steps as 0.1, 0.3, 0.5, and 0.9, respectively. The following type of variables are included - two continuous scaled variables, one binary scaled variable and one semi-continuous variable. As for all simulations in this section, the number of repetitions is 500, and the final error measure is the average of all 500 resulting error measures. The proportion of missing values is fixed with 5% in each variable.

The results of the algorithms `IVEWARE`, `IRMI`, and `IMI`, the non-robust version of `IRMI`, are presented in Figure 3. Generally, the error measure decreases with increasing correlation, because then the multivariate information is more and more useful for the estimation of the missing values. The error measure for the categorical variables (see Figure 3(a)) and the continuous and semi-continuous variables (Figure 3(b)) is compariable among the robust method (`IRMI`) and the non-robust OLS-based regression method, because no outliers have been generated in the simulated data. The difference for the error measures between `IVEWARE` and our proposed methods gets more pronounced with increasing correlation (Figure 3(b)). Here, `IVEWARE` obviously may suffer from a less optimal strategy to initialize the missing values, which is then reflected in slightly poorer imputation quality.



(a) Imputation error for binary and categorical parts

(b) Imputation error for (semi-)continuous parts

Figure 3: Comparison of the error measures resulting from the three algorithms by varying the correlation structure of the generated data.

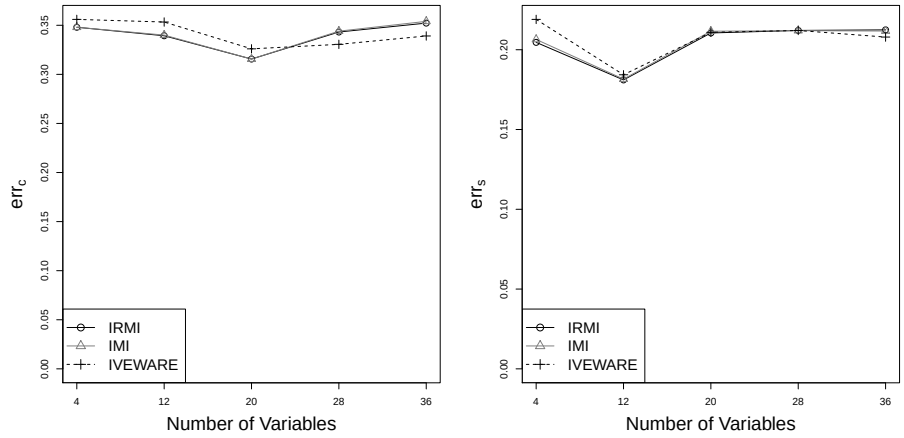
5.3 Second configuration: varying the number of variables

With increasing dimensionality of the data, the gain in multivariate information can be used by model-based regression imputation as long as the additional variables are not uncorrelated to the variable where missing information needs to be estimated. This effect is demonstrated by a simulation setting which starts from 4 variables (two continuous, one binary and one semi-continuous), and increases in each step the dimensionality of the data by including one further variable of each type. Figure 4 presents the results of this study. Here the simulated data are based on multivariate normally distributed data with fixed covariances of 0.7 and variances of 1. Again the proportion of missing values is fixed with 5% in each variable.

Figure 4 shows that all three algorithms have a similar performance. `IMI` and `IRMI` have a slightly better precision than `IVEWARE` with respect to lower dimensionality of the data. While the errors from the categorical and binary variables remain almost constant when increasing the number of variables (see Figure 4(a)), the error from imputing the continuous and semi-continuous scaled variables is first decreasing and then increasing (see Figure 4(b)).

5.4 Third/fourth configuration: varying the amount of outliers using variables with high/low correlation

To illustrate the influence of outliers on the considered imputation algorithms, n_1 out of n observations will be replaced by outlying observations. The non-outlying part of the data is generated in the same way as described in the previous settings, with covariances



(a) Imputation error for binary and categorical parts

(b) Imputation error for (semi-)continuous parts

Figure 4: Comparison of the error measures resulting from the three algorithms by varying the number of variables.

of 0.9 (third configuration) and 0.4 (fourth configuration), respectively, including two continuous scaled variables, two binary variables and one semi-continuous variable. The outlier part is generated with the mean vector

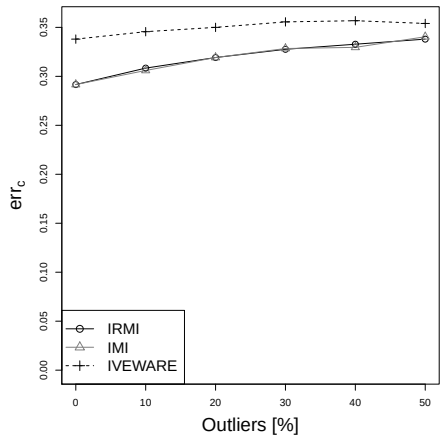
$$\mu_{out} = (5, 15, 10, 10, 10)^t$$

and covariances 0.5 (third configuration) and 0.4 (fourth configuration), the variances are 1. The generation of binary variables and semi-continuous variables is done as described in the first part of this chapter. The percentage of outliers is varied from 0 to 50. The proportion of missing values in the variables is $(0.1, 0.06, 0.05, 0.04)$, and they are only chosen in the non-outlying part. Accordingly, the error measures are only based on non-outlying observations. The results are shown in Figure 5.

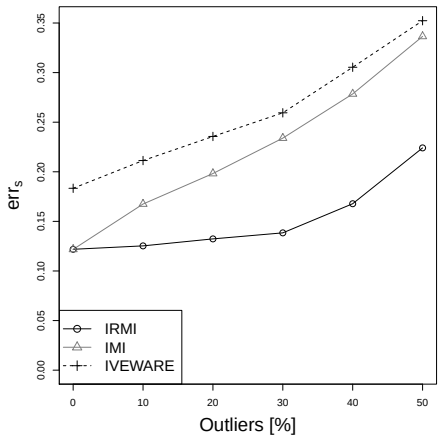
With respect to the errors in the categorical parts (Figure 5(a) and 5(c)), the non-robust method `IMI` and its robust counterpart `IRMI` performs almost identical since non-robust regression is used by `IRMI` for imputing categorical responses per default. Especially for the highly correlated data including outliers, both methods outperforms `IVEWARE`. The error in the continuous parts reveals a contrasting behavior (Figure 5(b) and 5(d)). Here, the robust version `IRMI` clearly dominates, and it remain stable until about 30% outliers. The non-robust version `IMI` is preferable over `IVEWARE`.

6 Application to real data

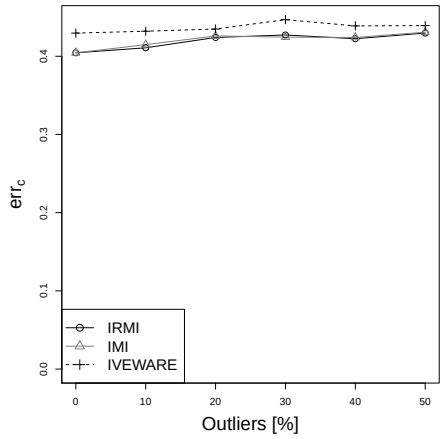
Two very popular complex data sets from official statistics are used to compare the imputation algorithms, one from social sciences (EU-SILC) and one from business statistics



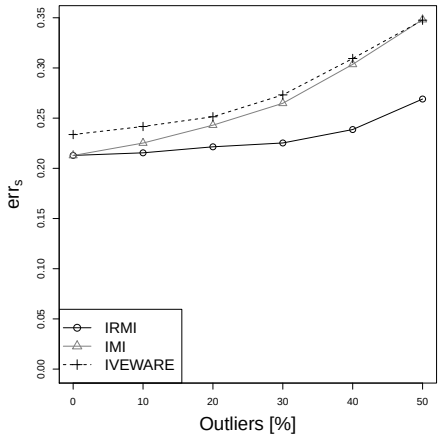
(a) Imputation error for binary and categorical parts; variables with high correlation



(b) Imputation error for (semi-)continuous parts; variables with high correlation



(c) Imputation error for binary and categorical parts; variables with low correlation



(d) Imputation error for (semi-)continuous parts; variables with low correlation

Figure 5: Comparison of the error measures resulting from the three algorithms by varying the percentage of outliers.

(SBS). The third real-world data set considered is a census data set. In addition to that, the airquality data set is considered, which is a popular data set often used in the literature about missing values.

Originally, all these data sets come with missing values. We took the available complete observations, almost ignoring the possible dependencies of the missing values in the data. However, also within this simplification the results should refect how the algorithm performs with the data. We set missing values randomly to the available information before imputing these artificial missing values. After imputation, the imputed values are compared with their "true" original values.

6.1 EU-SILC

To show results based on a complex real-world data set from official statistics, the European Union Statistics of Income and Living Conditions (EU-SILC) survey 2006 from Statistics Austria is chosen. This very popular data set is mainly used for measuring poverty and social inclusion in Europe. It includes a moderate amount of missing values in the semi-continuous part of the data. `IVEWARE` is used by various statistical agencies (e.g. by the Federal Statistical Office in Swiss) to impute the income components of the EU-SILC data, for example. Statistics Austria, for example, uses (non-iterative) least-squares regression imputation whereas a random error based on the variance of the residuals is added to the response [see, e.g., Rubin, 1987, Ghellini and Neri, 2004]. Fisher [2006] imputed the income components of "similar" data (consumer expenditure data) separetely, i.e. he uses one income compontent as the response variable and as predictors demographic characteristics of the consumer unit and a variable that equals the quarterly expenditure outlays for the consumer unit. He performs a stepwise backward approach to select only the most important predictors.

For imputation we used the household income variables (semi-continuous variables) with the largest amount of missing values in the raw data set, namely *hy050n* (family/children related allowances), *hy060n* (social exclusion not elsewhere classified), *hy070n* (housing allowances), *hy090n* (interest, dividends, profit from capital investments in unincorporated business), and three categorical variables, namely *household size*, *region*, and *number of childrens in the household*. The available complete observations of this data set are used (3808 observations), and missing values are set completely at random in the income components respecting the rate of missing values in the income variables from the complete data. Therefore, 25 percent missing values are generated in variable `hy090n`, and 2 percent missing values are generated in the other income components. `IMI`, `IRMI` and `IVEWARE` are applied to impute the missing values.

The procedure is repeated 1000 times. Figure 6 displays the results of the simulation. Since missing values are only obtained in the semi-continuous income components, only the error measure for continuous and semi-continuous variables is reasonable. It is easy to see that `IRMI` leads to much better results than `IMI` and `IVEWARE`. Also the variance of the errors is much smaller.

An additional result by imputing this data set with `IRMI` is obtained by Alfons et al. [2009]. They used `IRMI` (without describing the algorithm) to estimate the additional

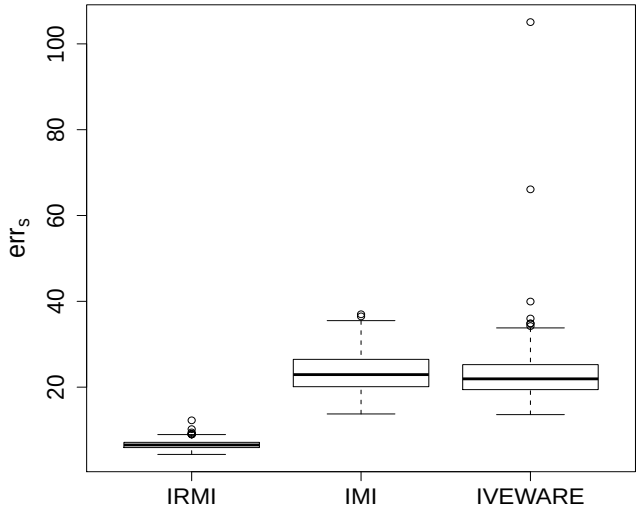


Figure 6: Results for the Austrian EU-SILC data comparing original data points with the imputed data points.

uncertainty (with respect to missing values) of indicators via the bootstrap approach from Little and Rubin [1987]. In fact they estimate the additional uncertainty due the presence of imputations when estimating the GINI coefficient but also the weighted mean of the equivalised household income from the Austrian EU-SILC data. The additional uncertainty was evaluated for the point estimates but also for the variance estimates. Their simulations was not designed to show that `IRMI` is proper according to definitions in Nielsen [2003] or Rubin [1987], but it results in realistic estimates and consider small additional uncertainty due to point and variance estimates.

6.2 Stuctural Business Statistics Data

The Austrian structural business statistics data (SBS) from 2006 covers NACE sections C-K for enterprizes with 20 or more employees (NACE C-F) or above a specified turnover (NACE G-K) [Eurostat, 2008b]. For these enterprizes more than 90 variables are available. Only limited administrative information is available for enterprizes below these thresholds. The raw unedited data consist of 21669 observations including 3891 missing values.

As mentioned in Section 1, imputation should be made in reasonable subgroups of the data. A detailed data analysis of the raw data has shown that homogeneous subgroups are based on NACE 4-digits level data. Broader categories imply that the data consist of different kinds of sub-populations with different characteristics. For the sake of this study

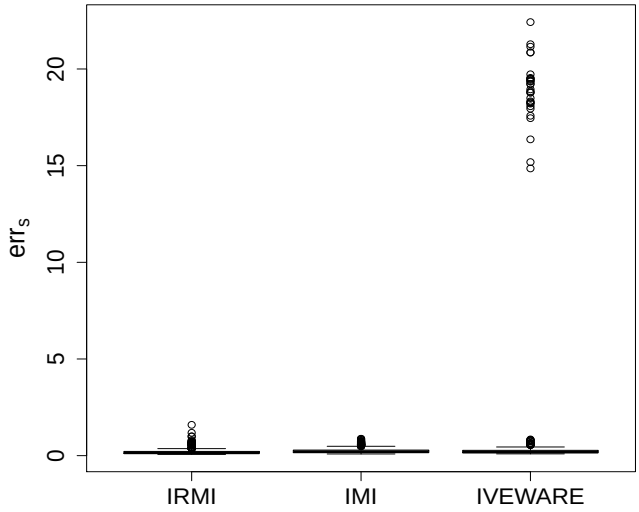


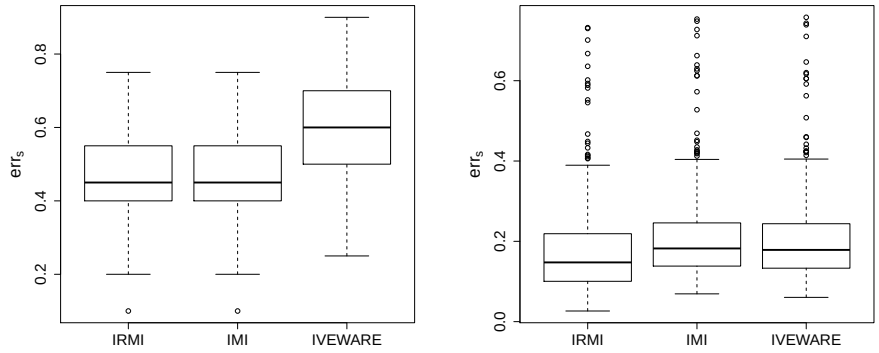
Figure 7: Results for the Austrian SBS data comparing original data points with the imputed data points.

we have chosen the NACE 4-digits level *47.71 - "Retail sale of clothing in specialized stores"* - (Nace Rev. 1.11 *52.42*, ISIC Rev.4 *4771*). This typical NACE 4-digits level data set consists of 199 observations with 7 missing values and various outliers. In order to be able to apply imputation methods reasonably, specific variables were chosen, namely *turnover* (continuous), *number of white-collar employees*, *number of blue-collar workers*, *part-time employees*, *number of employees* (all disrect variables but considered as continuous), *wages*, *salaries*, *supply of trade goods for resale*, *intermediate inputs* and *revenues from retail sales* (considered as continuous variables). We used the 192 complete observations of this data set. From those observations we set 5% of the values in variables *number of employees* and *intermediate inputs* to be missing completely at random (only those variables that include missing values in the original data set). Afterwards the missing values are imputed and the error rate given in Equation 6 is calculated. In Figure 7 the results of 1000 runs are visualized. `IRMI` outperforms the two other imputation methods. `IVEWARE` sometimes converges - driven by the outliers in the data - to a worst solution which is clearly reflected in Figure 7.

6.3 Census Data from UCI

A census data set from 1994 is used to evaluate both errors given in Equation 5 and 6. This data set is provided by the University of California. It is a data frame with 32561 observations on 14 variables. We used the complete observations of respondents which

belongs to the complete rows corresponding to native country 21 (610 observations) and variables *age*, *workclass*, *education*, *sex*, textitrelationship and *hours per week* (for details on the data, have a look at `http://www.ics.uci.edu/~mlearn/MLRepository.html`). Missing values are set in the variables *hours per week* and *relationship* (3% missing values each).



(a) Error in categorical variables

(b) Error in continuous variables

Figure 8: Results for the CENSUSN data comparing original data points with the imputed data points.

With respect to the errors in categorical variables shown in Figure 8(a), `IVEWARE` performs worst while `IMI` and `IRMI` leads to similar results. For the continuous and semi-continuous part, `IRMI` leads to higher precision as the other two non-robust methods.

6.4 Air Quality Data

This data are obtained by daily air quality measurements in New York, May to September 1973 [see also Chambers et al., 2008]. It consits of 154 observations on 6 variables (*ozone*, *solar*, *wind*, *temp*, *month*, *day*) whereas the first two variables contain missing values. The complete observations of the first 5 variables are used and missing values are set in the first two variables according to the rate of the original data set (37 and 7 values out of 111 complete observations).

Regarding to Figure 9, again `IRMI` provides the best results. However, the gain in precision to the other two methods is small since almost few moderate outliers can be found in this small data set.

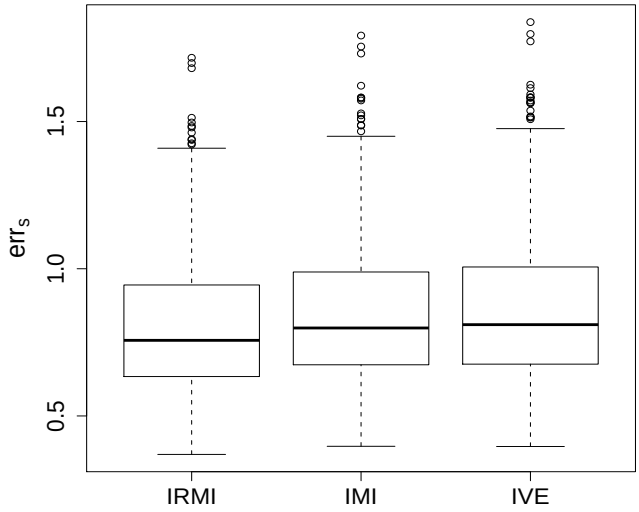


Figure 9: Results for the airquality data comparing original data points with the imputed data points.

7 Implementation

The algorithm `IRMI` (and the non-robust version `IMI`) has been implemented in the package `VIM` version 1.4 [Templ and Filzmoser, 2008, Templ et al., 2009] written in R [R Development Core Team, 2009] and freely distributed as open-source code via the web (see `http://cran.r-project.org`).

Let `X` be a data set already available in the R workspace, then

```
R> irmi(X, method='lmrob', mixed="VarMixed")
```

imputes the missing data. The variables of `X` can be of different scale (binary, nominal, ordinal, semi-continuous, continuous). The parameter `mixed` is a vector indicating the index of the position of the mixed scaled variables in the data (details are explained in the **VIM** manual [Templ et al., 2009]). The functon automatically detects binary and categorical variables, if the type is correctly assigned in R. It is possible to select either robust regression methods or standard OLS-based regression methods with the function parameter `methods` (details are described in the manual of **VIM**).

Print and summary methods as well as several diagnostic plots are provided to further analyze the results [see Templ et al., 2009]. For example, parallel coordinate plots and multiple scatterplots allow to highlight the imputed values in order to reveal the quality of the imputation.

8 Conclusions

All real-world data sets we have seen so far, especially in official statistics, include outlying observations and they often include different types of distributions. We proposed an iterative robust model-based imputation procedure for automatic imputation of missing values, which can deal with the mentioned data problems. All simulation results show that our robust method shows either equal behaviour or outperforms the investigated non-robust methods. Additionally, the results from the imputation of the complex and popular EU-SILC data set which includes several semi-continuous variables showed that `IRMI` performs very well in a real-world settings. Results from imputation of the SBS data, which is a very important data set in official statistics, go in the same direction and clearly shows that our proposed methods works better than non-robust methods like `IVEWARE`. Therefore, we would suggest to apply `IRMI` whenever an automatical approach for imputation is needed, and escpecially, when the variables are of mixed scale.

Furthermore, our methods are easily accessible since they can be freely downloaded from the comprehensive R archive network.

Acknowledgement

This work was partly funded by the European Union (represented by the European Commis- sion) within the 7th framework programme for research (Theme 8, Socio-Economic Sciences and Humanities, Project AMELI (Advanced Methodology for European Laeken Indicators), Grant Agreement No. 217322).

References

A. Alfons, M. Templ, and P. Filzmoser. On the influence of imputation methods on laeken indicators: Simulations and recommendations. In *UNECE Work Session on Statistical Data Editing; Neuchatel, Switzerland*, page 10, 2009. to appear.

C. Beguin and B. Hulliger. The BACON-EEM algorithm for multivariate outlier detection in incomplete survey data. *Survey Methodology*, 34(1):91–103, 2008.

E. Cantoni and E. Ronchetti. Robust inference for generalized linear models. *JASA*, 96 (455):1022–1030, 2001.

J.M. Chambers, W.S. Cleveland, B. Kleiner, and P.A. Tukey. *Graphical Methods for Data Analysis*. CA: Wadsworth, Belmont, 2008.

A.P. Dempster, N.M. Laird, and D.B. Rubin. Maximum likelihood for incomplete data via the EM algorithm (with discussions). *Journal of the Royal Statistical Society*, 39: 1–38, 1977.

G.B. Durrant. Imputation methods for handling item-nonresponse in the social sciences: a methodological review. Ncrm methods review papers, Southampton Statistical Sciences Research Institute (S3RI), University of Southampton, 2005.

Eurostat. Survey sampling reference guidelines. introduction to sample design and estimation techniques. Methodologies and working papers, isnn 1977-0375, European Commission, 2008a.

Eurostat. *NACE Rev. 2. Statistical classification of economic activites in the European Community*. Eurostat, Methodologies and Workingpapers, 2008b. ISBN 978-92-79-04741-1.

R.E. Fay. Alternative paradigms for the analysis of imputed survey data. *Journal of the American Statistical Association*, 91(434):490–498, 1996.

J.D. Fisher. Income imputation and the analysis of consumer expenditure data. *Monthly Labor Review*, 129(11):11–19, 2006.

G. Ghellini and L. Neri. Proper imputation of missing income data for the tuscany living condition survey. In *In Proceedings of the Atti della XLII Riunione Scientifica, Universitá di Bari, Italy*, page 4, 2004.

J. Honaker, G. King, and M. Blackwell. *Amelia: Amelia II: A Program for Missing Data*, 2009. URL http://CRAN.R-project.org/package=Amelia. R package version 1.2-2.

K. Hron, M. Templ, and P. Filzmoser. Imputation of compositional data using robust methods. Research report sm-2008-4, Department of Statistics and Probability Theory, Vienna University of Technology, 2008. URL http://www.statistik.tuwien.ac.at/forschung/SM/SM-2008-4complete.pdf.

P.J. Huber. *Robust Statistics*. Wiley, 1981.

R.J.A. Little and D.B. Rubin. *Statistical Analysis with Missing Data*. Wiley, New York, 1987.

Thomas Lumley. *mitools: Tools for multiple imputation of missing data*, 2010. URL http://CRAN.R-project.org/package=mitools. R package version 2.0.1.

R.A. Maronna, R.D. Martin, and V.J. Yohai. *Robust Statistics: Theory and Methods*. John Wiley & Sons, New York, 2006.

R. Muennich and S. Rässler. Variance estimation under multiple imputation. In *Proceedings of Q2004 European Conference on Quality in Survey Statistics, Mainz*, page 19, 2004.

S.F. Nielsen. Proper and improper multiple imputation. *Internat. Statist. Rev.*, 71(3): 593–607, 2003.

R Development Core Team. *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing, Vienna, Austria, 2009. URL http://www.R-project.org. ISBN 3-900051-07-0.

S. Raessler and R. Münnich. The impact of multiple imputation for DACSEIS. Research report ist-2000-26057-dacseis, 5/2004, University of Tübingen, 2004.

T.E. Raghunathan, J.M. Lepkowski, and J.V. Hoewyk. A multivariate technique for multiply imputing missing values using a sequence of regression models. *Survey Methodology*, 27(1):85–95, 2001.

B. D. Ripley. *Pattern Recognition and Neural Networks*. Cambridge University Press, 1996.

P.J. Rousseeuw and K. Van Driessen. Computing lts regression for large data sets. *Estadistica*, 54:163–190, 2002.

D.B. Rubin. *Multiple Imputation for Nonresponse in Surveys*. Wiley, New York, 1987.

J.L. Schafer. *mix: Estimation/multiple Imputation for Mixed Categorical and Continuous Data*, 2009. URL http://CRAN.R-project.org/package=mix. R package version 1.0-7, see also his implementation in SAS and SPLUS.

J.L. Schafer. *Analysis of Incomplete Multivariate Data*. Chapman & Hall, 1996. Chapter 9.

J.L. Schafer and M.K. Olson. Modeling and imputation of semicontinuous survey variables. Fcsm research conference papers, Federal Commitee on Statistical Methodology, 1999. URL http://www.fcsm.gov/99papers/shaffcsm.pdf.

Joseph L. Schafer and Maren K. Olsen. Multiple imputation for multivariate missing-data problems: a data analyst's perspective. *Multivariate Behavioral Research*, 33: 545–571, 1998.

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

M. Templ and P. Filzmoser. Visualization of missing values using the R-package VIM. Research report cs-2008-1, Department of Statistics and Probability Theory, Vienna University of Technology, 2008. URL http://www.statistik.tuwien.ac.at/forschung/CS/CS-2008-1complete.pdf.

M. Templ, A. Alfons, and A. Kowarik. *VIM: Visualization and Imputation of Missing Values*, 2009. URL http://cran.r-project.org. R package version 1.2.4.

S. van Buuren and C.G.M. Oudshoorn. Flexible multivariate imputation by MICE. Tno/vgz/pg 99.054, Netherlands Organization for Applied Scientific Research (TNO), 2005. URL http://web.inter.nl.net/users/S.van.Buuren/mi/docs/rapport99054.pdf.

W. N. Venables and B. D. Ripley. *Modern Applied Statistics with S*. Springer, 2002.

S. Verboven, K.V. Branden, and P. Goos. Sequential imputation for missing values. *Computational Biology and Chemistry*, 31:320–327, 2007.

V.J. Yohai. High breakdown-point and high efficiency estimates for regression. *The Annals of Statistics*, 15:642–665, 1987.

S. Yu-Sung, A. Gelman, J. Hill, and M. Yajima. Multiple imputation with diagnostics (mi) in R: Opening windows into the black box. *Journal of Statistical Software*, 2009. to appear.