Universität Hamburg
DER FORSCHUNG | DER LEHRE | DER BILDUNG

FB Psychologie
Psychologische Methoden

Discussion Paper

No. 4

Missing data: On criteria to evaluate imputation methods

January 2016

Daniel Salfrán, Pascal Jordan & Martin Spiess

Psychologische Methoden, Universität Hamburg, Von-Melle-Park 5, 20146 Hamburg,
Germany

Missing data: On criteria to evaluate imputation methods

Daniel Salfrán, Pascal Jordan and Martin Spiess*

Abstract

Empirical data analyses often require complete data sets. In the case of incompletely observed data sets therefore methods are attractive that generate plausible values (imputations) for the unobserved data. The idea is to then analyse the completed data set in an easy way, e.g. using publicly available software. Thus, various imputation methods have been proposed and evaluated. Popular measures used for evaluating these methods are based on distances between true and imputed values applied in simulation studies. In this paper we show through a theoretical example and a simulation study that these measures may be misleading: A small value of a measure which is a function of the distance between imputed and true values does not imply that inferences based on the imputed data set is somehow close to the (valid) inferences based on the complete data set without missing values. Hence, we propose to compare imputation methods with respect to their aptitude to enable valid inferences based on imputed data sets.

Keywords: Hit-rate criterion; Imputation techniques; Inference criterion; Mean squared error; Statistical properties.

[*] Corresponding author: Martin Spiess, Psychological Methods and Statistics, University of Hamburg, Von-Melle-Park 5, 20146 Hamburg, Germany; e-mail: martin.spiess@uni-hamburg.de

1 Introduction

Empirical data sets are often affected by missing data. Examples are microarray gene expression, other biological or medical laboratory or epidemiological data (e.g., Burgette and Reiter, 2010; Kshirsagar, Carbonell and Klein-Seetharaman, 2012; Moorthy, Mohamad and Deris, 2014; Nguyen, Wang and Carroll, 2004; Troyanskaya et al., 2001; Waljee et al., 2013), data from wireless sensor networks (e.g., Pan and Li, 2010) or traffic flow data (Li, Li and Li, 2014). Since downstream analyses usually require complete data sets, various imputation methods have been developed to replace unobserved by estimated values (imputations). Many of these methods are proposed based on comparisons using functions of squared distances between true and imputed values or the absolute values of these distances. Comparisons are usually carried out in simulation studies in which either data sets including missingness patterns are simulated or, given real data sets, only missing data are generated (internal validation; e.g., Burgette and Reiter, 2010; Kim, Golub and Park, 2005; Ingsrisawang and Potawee, 2012; Li, Li and Li, 2014; Malarvizhi and Thanamani, 2013; Moorthy, Mohamad and Deris, 2014; Myrtveit, Stensrud and Olsson, 2001; Nguyen, Wang and Carroll, 2004; Pan and Li, 2010; Sehgal, Gondal and Dooley, 2005; Troyanskaya et al., 2001; Waljee et al., 2013; Yoon, Lee and Park, 2007). Some of these methods are made available to researchers in public use software packages (e.g., Templ, Kowarik and Filzmoser, 2011).

In this paper we will show that these statistics to evaluate the performance may be misleading: From the fact that a measure is zero if all the imputed values are equal to the true but unobserved values, it does not follow that the smaller the value of such a measure, the 'closer' the inference based on the imputed data set to the (valid) inference based on the complete data set without missing values. Thus, a method characterized by a smaller value of such a measure is not necessarily superior to a method with a larger value with respect to the statistical properties of subsequent analyses. Moreover, these measures are usually calculated in simulations. But without further theoretical justification, the implications of corresponding studies using these measures are valid only for a very restricted number of situations: Statements about differences of true and imputed values in the realized situations.

The weakness of measures based on differences between true and imputed values will be shown by a theoretical counterexample in Section 4 and a simulation study using an imputation function whose usefulness has been justified by such a measure (`irmi`; Templ, Kowarik and Filzmoser, 2011) and which is available for the software package R (R Core Team, 2014) in Section 5. But first we will give a short introduction into the multiple imputation approach proposed by Rubin (1987) and discuss the mean squared error of imputations as a measure for the comparison of imputation methods in Section 2 and 3, respectively. Section 6 summarizes the results and proposes to adopt an evaluation criterion that focusses on the statistical properties of statistics in subsequent analyses.

2 Imputation

Replacing each missing value by just one estimated value and applying standard software may lead to severe problems. For example, if in subsequent analyses model parameters are estimated, then corresponding estimators may be biased if the method used to generate the imputations is not appropriate. If standard errors are calculated based on functions provided by the predominantly used software packages for completely observed data sets, they will systematically be downward biased even if the estimator of interest is (asymptotically) unbiased, leading to rejection rates which are too high and to confidence intervals which are too small (Rubin, 1987). Further, if the completed data set is used for generating predictions, then these predictions and/or their variances may systematically be biased. Thus the crucial question is about criteria that can be used to identify an appropriate imputation method.

In the context of multiple imputations, where for each missing value several ($m = 1, \ldots, M$) possible values — usually 5 to 20 — are generated, Rubin (1987) justified a Bayesian model based approach to generate imputations. However, non-Bayesian imputation methods, e.g., based on Bootstrap techniques may also be proper (Rubin, 1987). If an imputation method is proper, then estimators with desirable properties in complete data situations when applied to multiply imputed data sets can still be expected to have favourable properties, i.e., they should be consistent, asymptotically normally distributed and a variance estimator is available. Corresponding confidence intervals and hypothesis tests tend to be valid and valid predictions are usually possible.

Basically, an imputation method tends to be proper if the imputations are independent draws from an appropriate posterior predictive distribution of the variables with missing values given all other variables (Rubin, 1987). This does not only imply that the deviations of the imputed from the missing true values are unsystematic, but also that the variation in the imputations reflects all the uncertainty in the predictions. *It does neither require nor imply that some measure of distances between true and imputed values is minimal.* For more detailed discussions about proper imputation methods, see Rubin (1987, 1996, 2003), Little and Rubin (2002), Meng (1994), Meng and Romero (2003), Nielsen (1997, 2003a, 2003b) and Robins and Wang (2000).

Once an incomplete data set is multiply imputed, standard analyses are applied to the M data sets, and the results are easily combined according to the rules given in Rubin (1987): Let $\hat{\boldsymbol{\theta}}_m$ be an estimator of scientific interest based on the mth imputed data set ($m = 1, \ldots, M$), $\hat{\boldsymbol{\theta}}$ the mean of the M estimators and $\widehat{\mathrm{Var}}(\hat{\boldsymbol{\theta}}_m)$ the variance estimator of $\hat{\boldsymbol{\theta}}_m$, then the final estimator for $\boldsymbol{\theta}$ and its variance estimator are given by $\hat{\boldsymbol{\theta}} = \sum_m \hat{\boldsymbol{\theta}}_m / M$ and $\widehat{\mathrm{Var}}(\hat{\boldsymbol{\theta}}_M) = W_M + (1 + M^{-1}) B_M$, where $W_M = \sum_m \widehat{\mathrm{Var}}(\hat{\boldsymbol{\theta}}_m)/M$ is the within variability and $B_M = \sum_m (\hat{\boldsymbol{\theta}}_m - \hat{\boldsymbol{\theta}})(\hat{\boldsymbol{\theta}}_m - \hat{\boldsymbol{\theta}})'/(M-1)$ is the between variability, which reflects the amount of information due to nonresponse.

3 Evaluation Criteria

One strategy to evaluate an imputation method in the context of statistical analyses is actually to evaluate the method used to create the imputations with respect to the statistical properties of resulting estimators or, more generally, functions of the random variables considered and may be called 'inference criterion' (Kuchler and Spiess, 2009).

The intuitively appealing idea of evaluating imputation methods based on differences between imputed and true values can only be realized theoretically or in simulation studies. Since theoretical results are often not available in complex real world situations, simulations are usually used to demonstrate the performance of imputation methods (e.g., Waljee et al., 2013; Templ, Kowarik and Filzmoser, 2011). However, functions of distances of imputed and true values only consider just a few aspects of the predictive distribution of the variables to be imputed.

We will illustrate this based on the mean squared error (mse) of the imputations, noting that this measure has a slightly different meaning as the mean squared error of estimators used in statistics. Let X be a random variable and $\tilde{X}$ be another random variable used as prediction or imputation for X. For simplicity, we take expectations with respect to X and $\tilde{X}$ instead of means over units and simulations. Then the expected mean squared error (emse) of the imputations is

$$E[\text{mse}(\tilde{X})] = E[(\tilde{X} - X)^2] = E[\text{var}(\tilde{X}|X)] + E[(E(\tilde{X}|X) - X)^2].$$

Obviously, moments above the second are ignored in the evaluation criterion. However, in subsequent analyses also higher moments may have an effect on the properties of test statistics and estimators. For example, de Jong, van Buuren and Spiess (2014) found in simulations that imputation methods may work well according to bias and coverage of regression parameter estimators if the conditional distribution of variables to be imputed and the conditional distribution from which imputations were drawn are both symmetric. But the same methods may fail if the distribution of imputations conditional on all observed variable values differed in skewness or kurtosis from the true conditional distributions of the variables with missing values.

Further, $E[\text{mse}(\tilde{X})]$ is small if either $E[\text{var}(\tilde{X}|X)]$ or $E[(E(\tilde{X}|X)-X)^2]$ vanishes and the other term is small. However, from an inferential point of view it may be preferable to have an imputation method with zero bias of the imputations and to accept a higher variance. Yet, an imputation method with larger bias but small variance would be preferred based on the expected mean squared error. For example, assume that one imputation method would produce imputations with no variation but overestimating the unobserved true values X by a fixed amount a. The second imputation method is equal to the true data generating process and imputes with zero bias but with variance $b > a^2$. In this case, for example, the mean would systematically overestimate the true mean of X based on the first but not on the latter method.

Of course, it is desirable to have imputations with minimum variance. But a more important aspect is that the functions of observed and imputed variables, e.g. estimators or test statistics, and their variances can consistently be estimated. The requirement of a minimal variance of imputations given the true but unobserved variables may thus be in conflict with the idea underlying the method of multiple imputation that the variation in the imputations should reflect all the uncertainty in the predictions. In this latter situation, if the imputation method is proper, a simple variance estimator is available while in the former it is usually not (see, e.g., the variance estimator proposed by Robins and Wang, 2000). Hence if the imputation method is improper and the variation of the imputations does not reflect all the uncertainty in these predictions, then variance estimators in downstream analyses have to be adapted. Note that values of $[E(\tilde{X}|X) - X]$ equal or close to zero are implied by the requirement that the imputations are drawn from an appropriate predictive distribution, but the converse does not necessarily hold.

Note also that if the imputed values were identical to the true but unobserved values — and thus any measure based on differences were zero — then the resulting downstream analyses based on the complete data set without missing values and based on the completed data set were identical. But this happens only if either the imputations can be derived without error from other variables or with probability approaching zero if the imputation method is based on stochastic methods. The first case would usually not be called a missing data problem, and simple calculation rules could solve the problem. The latter case is a missing data situation, but there is no theoretical justification for the (implicit) assumption of a one-to-one relationship between the values of a function based on differences of true and imputed values and the properties of subsequent inferences. Hence, imputation methods which seem to be appropriate according to this 'hit-rate' criterion (Kuchler and Spiess, 2009) do not guarantee that downstream analyses are valid. Even more, if they are evaluated only with the help of simulations, statements beyond the situations considered in the study are not justified. The failure of the hit-rate criterion as general measure to evaluate imputation methods will be illustrated in the next section.

4 Counterexample 1: Binary Random Variables

Let X be a random binary variable and denote the unknown probability of observing a '1' as π, i.e. $\Pr(X = 1) = \pi$. Further assume that we independently repeat this Bernoulli process n times, so that $X_i \sim X$, $i = 1, \ldots, n$, $\Pr(X_i = 1) = \pi$ and $\Pr(X_i = 0) = 1 - \pi$ for all i.

Suppose further that not all n values of $X_1, \ldots, X_n$ are observed, and that for each i the probability of observing the value of X_i is independent from all X_i, $i = 1, \ldots, n$. If missing values are generated independently from all variables, the missing data are called missing completely at random (MCAR; Little and Rubin, 2002). For simplicity we rearrange the data such that out of the n results, the first n_{obs} are observed whereas the last n_{mis} are missing, hence $n = n_{\text{obs}} + n_{\text{mis}}$. Let $\sum_{\text{obs}} X_i$ denote summation over the first n_{obs} and $\sum_{\text{mis}} X_i$ denote summation over the last n_{mis} elements. Thus $\hat{\pi}_{\text{obs}} = \sum_{\text{obs}} X_i / n_{\text{obs}}$ is an unbiased estimator of π based on the observed values only.

We will consider two imputation strategies: Generate imputations deterministically based on the estimated probability using the observed part of the sample, i.e., if $\hat{\pi}_{\text{obs}} > 0.5$ impute the value '1', otherwise impute '0' for all missing values. This is an extension of an example considered in Rubin (1996) and will be denoted as strategy S1. Alternatively we may adopt an imputation model that tries to mimic the true data generating process following Rubin (1987). This strategy will be called strategy S2.

If S2 is adopted, then one could apply a Bootstrap technique to generate imputations as follows: 1) Select a Bootstrap sample, i.e. randomly select n_{obs} values with replacement from the observed part of the sample, 2) select for each missing value randomly (with replacement) a value from this Bootstrap sample. Repeating steps 1) and 2) M times generates M imputed versions of the data set.

Let $I(\text{A})$ denote the indicator function, taking on value one if event A occurs and zero otherwise. Imputations are denoted by $\tilde{X}$, thus $\tilde{X}_{i,m}$ denotes the mth imputed value of case i and is generated according to $\tilde{X}_{i,m} = I(\hat{\pi}_{\text{obs}} > 0.5)$. Index m will only be used if $m > 1$. Further, we will adopt a frequentist perspective and fix n and n_{obs} and thus n_{mis}.

Adopting S1 leads to

$$\begin{aligned}
\hat{\pi}_{S1} &= \Big(\sum_{\text{obs}} X_i + \sum_{\text{mis}} I(\hat{\pi}_{\text{obs}} > 0.5)\Big)/n, \\
E(\hat{\pi}_{S1}) &= (n_{\text{obs}}\pi + n_{\text{mis}}[1 - B])/n, && (1) \\
\text{var}(\hat{\pi}_{S1}) &= (n_{\text{obs}}\pi(1-\pi) + 2n_{\text{mis}}n_{\text{obs}}\pi(B - B_-) + n_{\text{mis}}^2(1-B)B)/n^2, && (2) \\
E(\text{mse}(\tilde{X})) &= E[(\tilde{X}_{i,m} - X_i)^2] = 1 - B + 2\pi B_- - \pi, && (3)
\end{aligned}$$

where $\lfloor k \rfloor$ is the greatest integer less than or equal to k, $B := B(\lfloor n_{\text{obs}}/2 \rfloor; n_{\text{obs}}, \pi)$ denotes the binomial distribution function evaluated at $\lfloor n_{\text{obs}}/2 \rfloor$, $B_- := B(\lfloor n_{\text{obs}}/2 - 1 \rfloor; n_{obs} - 1, \pi)$ and expectations are taken with respect to all X_i.

For S2 we first generate M imputations for each missing value of X and then calculate m versions of the estimator for π, $\hat{\pi}_m$, as described in Section 2. We get

$$\hat{\pi}_{S2} = \Big(\sum_{\text{obs}} X_i + [\sum_{\text{mis}} \sum_{m} \tilde{X}_{i,m}]/M \Big)/n,$$

$$E(\hat{\pi}_{S2}) = \pi \tag{4}$$

$$\text{var}(\hat{\pi}_{S2}) = \frac{\pi(1-\pi)}{n} \Big(1 + \frac{n_{\text{mis}}}{n_{\text{obs}}} + \frac{1}{M} \frac{n_{\text{mis}}}{n_{\text{obs}}} \frac{(n_{\text{obs}} - 1)(n-1)}{n_{\text{obs}} n} \Big) \tag{5}$$

$$E(\text{mse}(\tilde{X})) = E[(\tilde{X}_{i,m} - X_i)^2] = 2\pi(1-\pi). \tag{6}$$

To decide which of the two strategies, S1 or S2, should be preferred, one could first state that S1 is simpler, needs less computing time and is thus computationally more efficient than S2. Adopted criteria are often functions of the distances of true and imputed values (see, e.g., Moorthy, Mohamad and Deris, 2014), one version being the mean squared error of imputations, given as the mean of the squared differences of true an imputed values. In our example, the expected mse can be calculated under both strategies. The emse under S1, (3), is a function of π, B and B_-, where B is the probability for observing $\hat{\pi}_{\text{obs}} \leq 0.5$ given π and a fixed number n_{obs} of independent draws. B_- is the probability for observing $\hat{\pi}_{\text{obs}} \leq 0.5 - 1/n_{\text{obs}}$ in $n_{\text{obs}} - 1$ independent draws given π. On the other hand, emse under S2, (6), does only depend on π with its maximum at $\pi = 0.5$.

Figure 1 shows the emse's under both strategies for two different numbers of observed values, $n_{\text{obs}} = 50$ and $n_{\text{obs}} = 100000$. Obviously, emse under S1 is smaller or equal to emse under S2. It is equal under S1 and S2 only if $\pi = 0$, $\pi = 1$ or, for large n_{obs}, if $\pi = 0.5$. According to the emse, S1 would thus be preferred. Usually the comparison of the different imputation methods stops at this point (e.g., Troyanskaya et al., 2001; Waljee et al., 2013).

However, if the variance of the binary variable before any missing values occurred would be of interest, S1 would lead to a severe underestimation of the variability of X even for the actual data set. If more general downstream analyses are intended based on the imputed data sets, then additional aspects become relevant. For example, one might be interested in estimating π or the variance of the binary variable, testing hypotheses about π

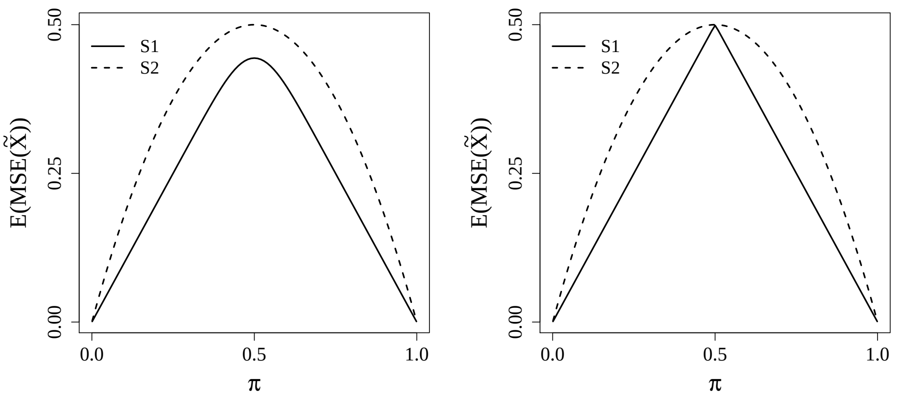

Figure 1: *emse under strategy S1 and S2 for* $n_{obs} = 50$ *(left) and* $n_{obs} = 100000$ *(right).*

or constructing confidence intervals. Then the statistical properties of $\hat{\pi}_{S1}$ and $\hat{\pi}_{S2}$ become important.

Result (1) shows that $\hat{\pi}_{S1}$ is unbiased only if $n_{\text{mis}} = 0$ or $B = 1 - \pi$. n_{mis} is zero, if there are no missing values, in which case $n_{\text{obs}} = n$. B is equal to $1 - \pi$, if $\pi = 0$ or $\pi = 1$. For large enough n, $\pi = 0.5$ is the only other point for which this function is approximately equal to $1 - \pi$. To see this, approximate the Binomial distribution by the normal distribution. Inspection of this function and its first and second derivatives with respect to π shows that it is equal to $1 - \pi$ if $\pi = 0.5$ and it is (strictly) concave for $0 < \pi < 0.5$ and (strictly) convex for $0.5 < \pi < 1$ with a negative first derivative for all π.

From (4) it is seen that $\hat{\pi}_{S2}$ is unbiased. Bias (bias), variance (var) and mean squared error (mse) of $\hat{\pi}_{S1}$ and $\hat{\pi}_{S2}$ for an example with $n_{\text{obs}} = 30$, $n_{\text{mis}} = 20$ and $M = 1$ are depicted in Figure 2. Note that the mean squared error of estimator $\hat{\pi}$ is $\text{mse}(\hat{\pi}) = E[(\hat{\pi} - \pi)^2] = [\text{bias}(\hat{\pi})]^2 + \text{var}(\hat{\pi})$.

Figure 2 shows that S1 leads to an estimator for π which is biased for almost every value of π and may have excessive variance in a region around $\pi = 0.5$. This leads to biased and/or invalid inferences over a wide range of π values. From the graph depicting the mse of both estimators, one may infer that there is a small range of values for which S1 leads to an estimator which, although biased, has a slightly lower mse. The smaller mse could make this estimator attractive in this area. It should be noted, however, that the

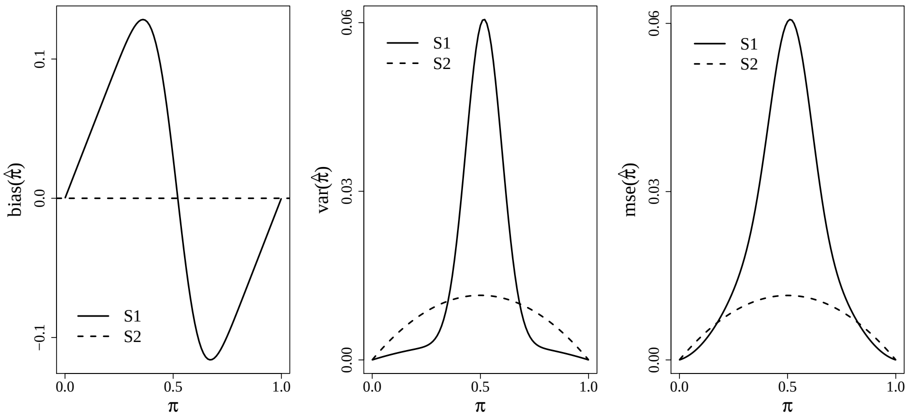

Figure 2: Bias*, variance* (var) *and* mse *of* $\hat{\pi}_{S1}$ *and* $\hat{\pi}_{S2}$ *for* $n_{obs} = 30$ *and* $n_{mis} = 20$*,* $M = 1$.

number of imputations can be increased under S2 leading to a decreasing variance of $\hat{\pi}_{S2}$, and eventually an mse which is not larger than that for $\hat{\pi}_{S1}$. In addition, increasing the number of observations n also leads to a disappearance of the small advantage of S1 over S2.

Thus, a small value of emse does not imply desirable properties of estimators of unknown model parameters. Further it is sample-specific (see Figure 1). In this simple example, the imputation strategy based on Rubin's (1987) theory has the larger emse, but leads to estimators with better properties over a wide range of π. Note that in general problems, variance estimators of estimators based on singly imputed data sets or based on data sets completed by adopting an improper imputation method (in the sense of Rubin, 1987), are often not available or at least very costly to get (c.f., Robins and Wang, 2000). But a simple variance estimator of $\hat{\pi}_{S2}$ based on $M > 1$ is available by the estimator given in Section 2. Further note that, in this example, considering the emse does not tell anything about the imputation methods beyond the function of the differences of imputed and true values: Even the variance of the binary variable in the imputed data set would be systematically too small as compared to the complete data set. Hence, the emse is a descriptive measure with a situation specific meaning. Any generalization to statements about the quality of an imputation method is not justified.

5 Counterexample 2: Continuous Random Variables

For the second example, we again adopted two different imputation methods. The first one is primarily justified in simulations by small values of an error measure which is a function of differences between imputed and true values. The method is realized in function `irmi` ('Iterative Robust Model-based Imputation'; Templ, Kowarik and Filzmoser, 2011) as part of the package `VIM`, available for R (R Core Team, 2014). `irmi` is a robust imputation function, but a non-robust version is available by choosing the corresponding option.

The other method is implemented as `mice` (van Buuren et al., 2014), a function also available as an R package, and — if only one variable is affected by missing values — allows generating imputations from a predictive posterior distribution (Rubin, 1987) based on a linear regression model for continuous variables, which tries to mimic the corresponding data generating process in its relevant aspects. It is written to handle much more general situations, but the situation considered here is a special case. Both methods allow the generation of multiple imputations, and we use $M = 10$ throughout. Since in our simulation study we want to compare imputation methods that differ mainly in their general justification, we use the parametric version of `irmi` to compare it with the parametric imputation method implemented in `mice`.

We simulated data sets with four continuous variables varying the following conditions:

1. number of units n, with $n = 200$ or $n = 2000$,

2. distribution of variables, with a) variables multivariate normally distributed ('normal' condition) and b) one t-distributed ($df = 3$), one χ^2-distributed ($df = 4$), one uniformly on $[0, 1]$ distributed variable — these three variables were independently generated — and the fourth variable being a weighted linear combination of the former variables with weights $\boldsymbol{\beta} = (-2, -1, 4)'$ plus a normally distributed error term with zero mean and variance one ('mixed' condition),

   (a) under the normal condition: the covariance matrix of the variables was a correlation matrix with all correlations being equal to either $\rho = 0.2$ or $\rho = 0.7$, the mean of each variable was 10, missing values are missing at random (MAR, see below),

(b) under the mixed condition: the missing data were either MAR or MCAR.

After having generated the data, we estimate a standard linear regression model of the fourth variable, denoted as Y, on all the others (X_1, X_2 and X_3) and a constant. To save space, we consider only the estimated parameter values weighting the third variable (X_3) in our data set. In case of the normal condition, the true parameter values of the corresponding regression weights depend on the mean, variances and the correlations of the variables. For $\rho = 0.2$ the corresponding weight of the third variable is $\beta = 0.143$ and for $\rho = 0.7$ we have $\beta = 0.292$. In the mixed condition $\beta = 4$.

Estimation of the regression parameters before any values are deleted serves as a comparison, results are labeled 'com' in Table 1. In a next step, missing values were generated. Under all conditions we generated approximately 40% missing values in the third variable. No other variable was affected. For missing values to be MCAR under the mixed condition, we generated binary response variables from a Bernoulli distribution with probability 0.6 and set values of X_3 to 'missing' if this variable was equal to zero, hence the missing probability was 0.4. For missing values to be MAR, we generated a latent variable as a linear function of Y and generated again binary response variables but now, under the normal condition, with probability 0.9 if the value of the latent variable was smaller than its median, and with 0.3 otherwise. In this case, the overall missing probability was again $0.4 = 1 - (0.9 \times 0.5 + 0.3 \times 0.5)$. Under the mixed condition, the response probabilities were 0.9 and 0.15 depending on whether the values of the latent variable was below its 0.6-quantile or not, and thus the overall probability of a missing value was again 0.4.

After values have been deleted, for each missing value, $M = 10$ imputations are generated with `irmi` (package `VIM` version 7.4; option for 'noisy' imputations turned on to generate variation in the imputations; see Templ, Kowarik and Filzmoser, 2011) and with `mice` (version 2.22). In both cases, finally a linear regression model is estimated using the combining rules given in Section 2. Note that differently to example 1, in example 2 the number of observed and missing values is not fixed.

For each of the eight conditions, we generated 1000 data sets and calculated the mean of the estimates (mean), the proportion of cases for which the confidence interval covered the true value (cover) and the mean of the values of

$$\text{err} = \frac{1}{Mn_{\text{mis}}} \sum_{m=1}^{M} \sum_{\text{mis}} |\tilde{x}_{i,m} - x_i| / |x_i|$$

(c.f. Templ, Kowarik and Filzmoser, 2011), but now taking into account that $M > 1$. In addition, we calculated the normalized root mean squared error (nrmse; e.g., Moorthy, Mohamad and Deris, 2014),

$$\text{nrmse} = \sqrt{\frac{\sum_{m=1}^{M} \sum_{\text{mis}} (\tilde{x}_{i,m} - x_i)^2}{M \sum_{\text{mis}} x_i^2}}.$$

Each simulation can be considered to be a binary experiment in which the confidence interval either covers the true value or not. Therefore, if all assumptions are met, the actual coverage rate based on 1000 replications should approximately be between 0.936 and 0.964 for $\alpha = 0.05$. Results are presented in Table 1.

Table 1: *Mean of estimates* (mean), *coverage* (cover; $\alpha = 0.05$) *and error measure* (err); *1000 simulations, 10 imputations.*

normal condition, MAR, $\rho = 0.2$, $\beta = 0.143$

	$n = 200$			$n = 2000$		
	mean	cover	err	mean	cover	err
com	0.145	0.958	0	0.143	0.945	0
irmi	0.148	0.897	0.104	0.151	0.865	0.104
mice	0.142	0.933	0.111	0.143	0.940	0.108

normal condition, MAR, $\rho = 0.7$, $\beta = 0.292$

	$n = 200$			$n = 2000$		
	mean	cover	err	mean	cover	err
com	0.294	0.960	0	0.292	0.941	0
irmi	0.306	0.914	0.066	0.308	0.836	0.065
mice	0.291	0.951	0.070	0.291	0.935	0.066

mixed condition, MCAR, $\beta = 4$

	$n = 200$			$n = 2000$		
	mean	cover	err	mean	cover	err
com	4.000	0.944	0	4.000	0.953	0
irmi	4.036	0.902	2.162	4.019	0.849	2.253
mice	4.000	0.955	2.312	4.001	0.947	2.395

mixed condition, MAR, $\beta = 4$

	$n = 200$			$n = 2000$		
	mean	cover	err	mean	cover	err
com	3.999	0.956	0	4.001	0.964	0
irmi	4.206	0.464	2.660	4.043	0.605	1.947
mice	4.009	0.950	2.452	4.010	0.944	2.233

The results in Table 1 suggest that the statistical properties of the downstream analysis based on imputations generated with `irmi` are systematically worse as compared to those based on mice: Estimates based on `irmi` show a systematic and larger bias than those based on `mice` and coverage rates are systematically too small. Coverage rates for results based on `mice` are in general acceptable with two borderline cases.

These results are in contrast to the mean values of the error measure: In all but one case, the error measure takes on smaller values for the imputations based on `irmi`, the exception being the mixed condition, where missing values are MAR and $n = 200$, in which case the value is smaller for imputations generated with `mice`. The pattern is the same if instead of err the measure nrmse is considered (not shown).

Thus, the values of the measures considered based on distances of imputed and true values do not reflect the statistical properties of downstream analyses. As such, they may be misleading: In our example, based on err or nrmse, one would prefer an imputation method that leads to biased estimators and coverage rates which can expected to be severely downward biased, leading to falsely rejecting a true null hypotheses much too often.

6 Conluding remarks

In this paper we considered popular measures to evaluate imputation methods in simulations which are based on a hit-rate criterion. These measures are functions of distances between imputed and true values. Adopting these mse-like measures presupposes that a method which implies values close to zero may be better in some sense than a method that leads to larger values. However, these measures lack a theoretical justification which would allow a general evaluation of imputation methods and would necessarily include a definition of the criteria of interest.

Of course one could be interested only in predictions given the specific simulated situation without requirement for more general inferences. In this case, one could define a function of distances of true and imputed values as a quality measure, run simulations for this situation and could infer from the results of these simulations which of the imputation methods seems to be optimal in the considered situations with respect to the specific criterion formalized by this measure. Without further justification, more general inferences would not be possible.

The chosen imputation method may not be optimal even according to the same criterion realized by this measure in a different situation. It may not allow to choose the best method if in the same or in other situations another aspect is of interest, like the variation of variables. And, as has been shown in the examples above, it fails if statistical properties of estimators and test statistics in downstream analyses are of interest. This is because there is no one-to-one relationship between measures based on differences of imputed and unobserved true values and properties of statistics considered to be important in subsequent analyses.

The problem is that these measures simply tell us nothing about other criteria than the distances in the actual data set. Even worse, as the true but unobserved values are not available in real-life situations and the true situation is unknown, simulation results based on a mse-like measure can not be used to justify the use of the corresponding imputation method in real-life applications. In this sense these measures are descriptive, situation-specific measures.

If more general statements are intended based on imputed data sets, other criteria need to be adopted. The criterion needed to preserve statistical validity in very general settings is the inference criterion. In the case of evaluating imputation methods, one way — although not the only one (e.g., Robins and Wang, 2000) — to realize this criterion is to generate multiple imputations according to proper imputation methods as described in Rubin (1987, 1996). Although it is not easy to develop methods that achieve this goal in general settings, it seems to be worth to develop them for allowing valid downstream inferences. If a theoretical justification for the use of measures to evaluate imputations methods is not available, simulation studies can be helpful. But instead of functions of distances between imputed and true but unobserved values, evaluation of imputation methods should rather focus on the properties of statistics, like the (asymptotic) bias of the estimators, the (asymptotic) bias of variance estimators and the actual coverage rates of corresponding confidence intervals.

References

Ingsrisawang, L., and Potawee, D. (2012). Multiple Imputation for Missing Data in Repeated Measurements Using MCMC and Copulas. *Proceedings of the International MultiConference of Engineers and Computer Scientists, Vol II. IMECS 2012,* March 14–16, Hong Kong.

Burgette, L. F., and Reiter, J. P. (2010). Multiple Imputation for Missing Data via Sequential Regression Trees. *American Journal of Epidemiology, 172,* 1070–1076.

de Jong, R., van Buuren, S., and Spiess, M. (2014). Multiple imputation of predictor variables using generalized additive models. *Communications in Statistics – Computation and Simulation.* doi: 10.1080/03610918.2014.911894.

Kim, H., Golub, G. H., and Park, H. (2005). Missing value estimation for DNA microarray gene expression data: local least squares imputation. *Bioinformatics, 21,* 187–198. doi:10.1093/bioinformatics/bth499

Kshirsagar, M., Carbonell, J., and Klein-Seetharaman, J. (2012). Techniques to cope with missing data in host-pathogen protein interaction prediction. *Bioinformatics, 28 ECCB,* 466–472. doi:10.1093/bioinformatics/bts375

Kuchler, C., and Spiess, M. (2009). The Data Quality Concept of Accuracy in the Context of Public Use Data Sets. *AStA Wirtschafts- und Sozialstatistisches Archiv, 3,* 67–80.

Li, Y., Li, Z., and Li, L. (2014). Missing traffic data: comparison of imputation methods. *IET Intell. Transp. Syst., 8(1),* 51–57.

Little, R. J. A., and Rubin, D.B. (2002). *Statistical analysis with missing data, (2nd ed.).* New York, John Wiley.

Malarvizhi, R., and Thanamani, A. S. (2013). Comparison of Imputation Techniques after Classifying the Dataset Using Knn Classifier for the Imputation of Missing Data. *International Journal Of Computational Engineering Research (ijceronline.com), 3(1),* 101–104.

Meng, X.-L. (1994). Multiple-Imputation Inferences with Uncongenial Sources of Input. *Statistical Science, 9,* 538–558.

Meng, X.-L., and Romero, M. (2003). Discussion: Efficiency and Self-efficiency With Multiple Imputation Inference. *International Statistical Review, 71,* 607–618.

Moorthy, K., Mohamad, M. S., and Deris, S. (2014). A Review on Missing Value Imputation Algorithms for Microarray Gene Expression Data. *Current Bioinformatics, 9,* 18–22.

Myrtveit, I., Stensrud, E., and Olsson, U.H. (2001). Analyzing Data Sets with Missing Data: An Empirical Evaluation of Imputation Methods and Likelihood-Based Methods. *IEEE Transactions On Software Engineering, 27(11),* 999–1013.

Nguyen, D.V., Wang, N., and Carroll, R.J. (2004). Evaluation of Missing Value Estimation for Microarray Data. *Journal of Data Science, 2,* 347–370.

Nielsen, S. (1997). Inference and missing data: Asymptotic results. *Scandinavian Journal of Statistics, 24,* 261–274.

Nielsen, S. (2003a). Proper and improper multiple imputation (with discussion). *International Statistical Review, 71,* 593–627.

Nielsen, S. (2003b). Rejoinder. *International Statistical Review, 71,* 625–627.

Pan, L., and Li, J. (2010). K-Nearest Neighbor Based Missing Data Estimation Algorithm in Wireless Sensor Networks. *Wireless Sensor Network, 2,* 115–122.

R Core Team: R (2014). *A Language and Environment for Statistical Computing.* R Foundation for Statistical Computing: Vienna, Austria. (http://www.R-project.org)

Robins, J. M., and Wang, N. (2000). Inference for Imputation Estimators. *Biometrika, 87,* 113–124.

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

Rubin, D. B. (1996). Multiple Imputation After 18+ Years. *Journal of the American Statistical Association, 91,* 473–489.

Rubin, D. B. (2003). Discussion on Multiple Imputation. *International Statistical Review, 71,* 619–625.

Sehgal, M. S. B., Gondal, I., and Dooley, L. (2005). A collateral missing value estimation algorithm for DNA microarrays. *IEEE International Conference on Accoustics, Speech and Signal Processing (ICASSP 05)*, 18–23, Philadelphia.

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

Troyanskaya, O., Cantor, M., Sherlock, G., Brown, P., Hastie, T., Tibshirani, R., Botstein, D., and Altman, R. B. (2001). Missing value estimation methods for DNA microarrays. *Bioinformatics, 17,* 520–525.

van Buuren, S., Groothuis-Oudshoorn, K., Robitzsch, A., Vink, G., Doove, L., and Jolani, S. (2014). *Package 'mice'.* Version 2.21. http://www.stefvanbuuren.nl

Waljee, A. K., Mukherjee, A., Singal, A. G., Zhang, Y., Warren, J., Balis, U., Marrero, J., Zhu, J., and Higgins, P.D. (2013). Comparison of imputation methods for missing laboratory data in medicine. *BMJ Open, 2013;3:e002847.* doi:10.1136/bmjopen-2013-002847

Yoon, D., Lee, E.-K., and Park, T. (2007). Robust imputation method for missing values in microarray data. *BMC Bioinformatics, 8(Suppl 2):S6.* doi:10.1186/1471-2105-8-S2-S6

Current Titles Discussion Papers FB Psychologie, Psychologische Methoden

1 de Jong, R., van Buuren, S. & Spiess, M. (2013). *Multiple imputation of predictor variables using generalized additive models.*

2 Jordan, P. & Spiess, M. (2013). *Fundamentale Probleme beim Einsatz testtheoretischer Modelle zur Diagnose von Individuen.*

3 Salfran, D. & Spiess, M. (2015). *A Comparison of Multiple Imputation Techniques.*

4 Salfrán, D., Jordan, P. & Spiess, M. (2016). *Missing data: On criteria to evaluate imputation methods.*