Towards a standardized evaluation of multiple imputation routines

Hanne Oberman and Gerko Vink

19-1-2022

Abstract

Developing new imputation methodology has become a very active field. Unfortunately, there is no consensus on how to perform simulation studies to evaluate the properties of imputation methods. In this paper we propose a move towards a standardized evaluation of imputation methods. To demonstrate the need for standardization, we highlight a set of potential pitfalls that bring forth a chain of potential problems in the objective assessment of the performance of imputation routines. This may lead to sub-optimal use of multiple imputation in practice. Additionally, we suggest a course of action for simulating and evaluating missing data problems.

Introduction

Multiple imputation (Donald B. Rubin 1987) is a state-of-the-art technique for drawing valid conclusions from incomplete data. The technique has earned a permanent spot in research and policy making, demonstrated e.g. by the detailed manual created by the National Research Council (Roderick J. Little et al. 2012). Although a top-down enforcement of valid ways to handle missing data is not yet very pronounced, an increasing amount of researchers are embracing multiple imputation techniques. After all, the principle of multiple imputation is very intuitive.

The idea behind multiple imputation is to impute (fill in) the missing values in incomplete data multiple times, to obtain a valid estimate of what could have been observed. The multiple datasets that are thus obtained can be analyzed by standard techniques and the analysis results can be combined into a single inference. In contrast to ad hoc methods for dealing with missing values (e.g. list-wise deletion, mean imputation, last observation carried forward), multiple imputation properly takes the sources of uncertainty that are related to the missingness problem into account. [It has been empirically demonstrated time and again that the inferences on multiply imputed data yield unbiased and confidence valid results (through Rubin's rules). But how do we know that this is the case, if we do not have access to the true but unobserved values of the missing entries? And how can we be sure that extensions to the methodology yield valid inferences too? Especially in the field of ML/AI, there are promising new imputation methods, which may yield even sharper imputations than now-standard (semi-)parametric imputation methods. That's why we use simulation studies for the evaluation of imputation routines.]

~~The quality of a solution obtained by multiple imputation depends on the statistical properties of the incomplete data and the degree to which an imputation procedure is able to capture these properties when modeling missing values. In general it holds that modeling missing data becomes more challenging when the amount of missingness increases. However, when (strong) relations in the data are present, the observed parts can hold great predictive power for the models that estimate the missingness. In that case, multiple imputation would be substantially more efficient than the ubiquitous list-wise deletion.~~

When evaluating the statistical properties (and thereby the practical applicability) of imputation methodology, researchers most often make use of simulation studies. In such studies a complete dataset is usually generated from a statistical model, another model is used to induce missingness, [then several missing data methods are applied] and a set of evaluation criteria is postulated to evaluate the performance of the imputation methods. However, no golden standard has been established to evaluate these imputation routines

and, as a result, the validity of the simulation procedures may differ tremendously from one developer to another.

The purpose of this paper is threefold: First, to raise some concerns with respect to evaluating imputation methodology. These concerns stem from careful consideration with fellow 'imputers' and from encounters as a reviewer for statistical journals. Second, to provide imputation methodologists with a suggested course of action when simulating missing data problems. This suggested approach should identify a common ground, but is in no way intended as an absolute solution. This identifies the third purpose of this paper: discussion. We hope to elicit critical thinking with respect to the problems at hand. We are all convinced that our methodology has some merit. But for sake of progress it would be much more advantageous if the aim of our evaluations would go beyond *proving the point* and would legitimately consider the statistical properties.

Why some evaluations should not be trusted

As of today, there is no consensus on how to perform simulation studies to evaluate the properties of imputation methods. Typically, the developer of a new imputation routine does some tests by simulations, but these tests differ across developers. This brings forth a chain of potential problems in the objective assessment of the performance of imputation routines that may lead to sub-optimal use of multiple imputation in practice. To demonstrate the broad impact of these problems, we subdivide the problems in the following three distinct categories: problems with data generation, problems with missingness generation and problems with performance evaluation. we further detail the impact these problems may have on the validity of the performance evaluation of the imputation routines.

Data generation problems

To evaluate the ability of an imputation routine to handle ~~the non-response model~~ [missing data entries], a form of truth has to be established. Those who perform simulation studies are in the luxury position to establish the truth beforehand, by choosing a data generating mechanism to obtain a complete dataset. Complete data is typically either generated from a statistical model, or sampled from a sufficiently large ~~observed~~ set [model-based vs design-based simulation]. ~~However, when generating data, two easily overlooked problems may arise.~~

The problem with model-based data generating mechanisms is that method performance may not translate to empirical data. First, data are often generated following the model (or distribution) that is also used for imputing the data. The performance of the imputation approach is then deemed good, which is no surprise as the evaluated conditions are in favor of the problem that is studied (Morris, White, and Crowther 2019). Although such procedures may be statistically relevant, the approach would be no good for even the simplest empirical case. Such simulations pose serious threats to the applicability of the method as real-life data hardly ever follow a given theoretical distribution. [Add that data types may be more messy.] Second, data are often generated such that the problem that is being studied is most pronounced. This results in simulated data that contains very valuable information structures, i.e. the correlations between groups of variables may be very pronounced. In other words, no matter what type of missingness is induced, the observed parts of the data may still hold much, if not all, of the information about the missing part and the performance of the imputation procedure is not surprisingly evaluated as very good. Also, it may be that the relations in real-life data are rarely as pronounced as the problem that is being studied.

The problem with design-based simulation is that it's difficult to find large datasets without missing entries. And imputing the missing data beforehand and then treating the once-imputed data as comparative truth may favor the used imputation method in further evaluations. A second issue is that SEs may be incorrect if the (re)sampling is not handled correctly. [Add pooling when the sample is the population (Vink and van Buuren 2014) and write something about bootstrapping SEs].

Problems with either of these data generation approaches is that the number of observations, the number of variables, the variable type(s) and the coherence between variables may give some imputation methods a comparative advantage, if they are not a) varied across simulation conditions [when is this fair? is this even

an option?], or b) the object of the study [add some examples of simulations specifically focused on small samples, categorical data, etc.].

[Maybe add that some simulation studies do not use a 'ground truth' at all, they just look for the method with the best predictive performance for a certain completely observed target variable. I don't know where to introduce the difference between statistical inference and prediction. But this is an important third category, it seems. These studies apply different combinations of imputation and estimation methods on one or more benchmark datasets, and assess the RMSE of the predicted values after imputation and estimation to say something about the comparative performance of the methods. In this context, there is no talk of retrieving the 'true' but unobserved values at all. Refer to missing data chapter (Liu et al. 2021).]

Missingness generation problems

Missingness comes in many shapes and forms [platitude, but how else to start?]. The most obvious characteristics of missing data are the missingness pattern(s) and missingness mechanism(s). Missingness patterns concern the location of missing entries across variables in an incomplete dataset, whereas missingness mechanisms describe the relationship between missingness and the values of variables in the data (Little and Rubin 2020, 8).

The missing data pattern may inadvertently guide which imputation method prevails in simulation studies. Some ad hoc methods are known to yield valid inferences in very specific missing data patterns (e.g., listwise deletion when there is missingness in the outcome variable of the analysis model exclusively), while they bias inferences in any other case. Extrapolating simulation results from such a specific missing data pattern to more intricate empirical missingness patterns could lead to false conclusions [re-read this! and add monotone/univariate vs multivariate and sporadic vs systematic dichotomies]. Related to the missingness pattern across variables is the missingness proportion. The terminology on this topic can be confusing. Does a missingness proportion of 50% mean that half of the entries in an incomplete dataset are missing, or that half of the rows have at least one missing entry? We will refer to the latter as the proportion of incomplete cases. But this is not a clear distinction in the literature, which may lead to incorrect conclusions [re-phrase this too]. Moreover, the amount of missingness that is used to simulate the performance of imputation methods differs between studies. [Add ref to chapter (Liu et al. 2021) and add info.] Some studies use only 10% missing data where other studies push the limits to additionally investigate the performance under at least 50% missing data. This inconsistent display of simulation results may impact the objectivity of meta-evaluations over imputation methods, as one method's performance may appear to be favorable because of the less stringent simulation conditions. This ultimately may lead to statisticians recommending a less efficient method to applied researchers, thereby limiting the efficiency of the imputation approach and unnecessarily lowering the statistical power.

[Briefly introduce missingness mechanisms here and show potential problems, but without simulation results (moved to separate section). This part should contain: 1) definitions of mechanisms, 2) that MCAR is often ignored, while it's the minimum requirement and sometimes realistic in practice, 3) that MAR can be generated *spuriously* if the correlation between variables is low, and 4) that MNAR is often ignored, while it may be the most realistic in practice. And finally something about types of M(N)AR, but what are specific problems with that?]

[Until next hline: to be edited!] In order to obtain valid imputation inference, the imputation model must capture the essence of the true non-response mechanism Meng (1994). The model - if any - that is used to generate the missingness is usually assumed to be random (MAR) or completely random (MCAR). Missingness mechanisms that are not random (MNAR) and mechanisms that are considered non-ignorable see e.g. Donald B. Rubin (1976), are generally ignored during evaluation of imputation methodologies, except for those methods that are specifically targeted at non-ignorable applications.

With MCAR missingness mechanisms, the probability to be missing is the same for all cases. This is a necessary simulation condition for evaluating the performance of imputation procedures. If an imputation method is not able to solve the problem (i.e. yield valid inference) under MCAR, the statistical properties of the procedure are not sound. Sadly, the straightforward case of MCAR is often neglected from simulation studies and focus is drawn to the evaluation of MAR mechanisms only. Alternatively, some studies limit their evaluations to MCAR mechanisms only, which in my view may be far too simplistic.
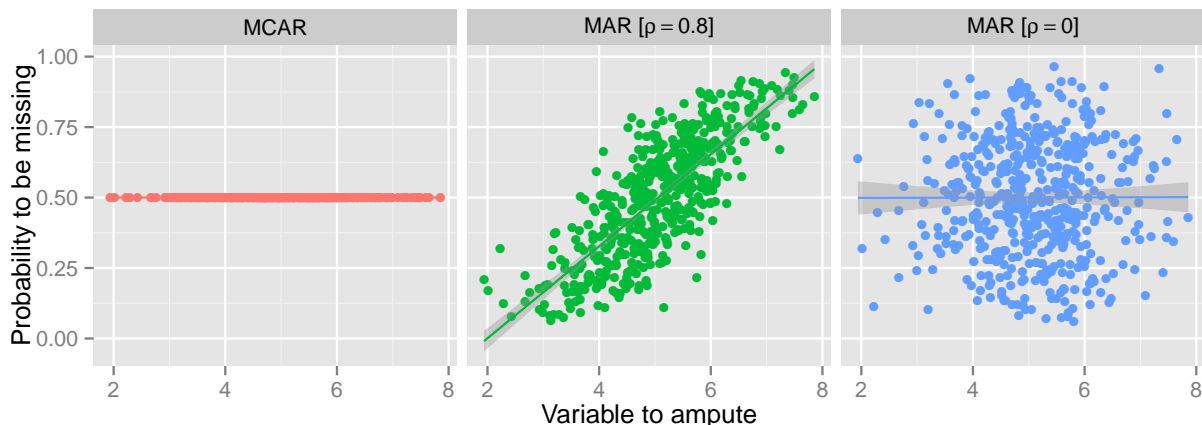

Figure 1: Three missingness mechanisms that yield approximately 50 percent missingness ($N = 500$). Displayed are a MCAR mechanism, a right-tailed MAR mechanism generated from bivariate normal data with correlation $\rho = .8$ and the same MAR mechanism, now generated from bivariate normal data with correlation $\rho = 0$.

With MAR missingness mechanisms, the probability to be missing is the same within groups of cases defined by the observed data only (e.g. males are less likely to disclose their weight, but gender is observed). The essence here is that the observed relations in the data are used to induce missingness during simulation (e.g. weight is made incomplete based on gender), but these relations may be weak, or non-existent.

When MAR is induced based on weaker relations in the data, claims for a method's applicability to situations where the missingness is random become less valid. When MAR is induced from data without multivariate relations, the inferential implication of the missingness would mimic those of MCAR. This problem is generally overlooked in simulation studies and the procedure to generate missing data is often insufficiently described in resulting publications. Figure 1 demonstrates two realizations of missingness mechanisms under MAR mechanisms derived from data with different multivariate relations.

It is obvious that the MAR missingness generated from data with low correlation would yield similar inference to the MCAR missingness, even though the missingness is clearly random. The simulation results from Table 1 confirm this. A detailed simulation setup can be found in Appendix A.

As expected, under genuine MAR missingness, i.e. the missingness probabilities differ conditional on the level of the incomplete variable, the performance of complete case analyses drops dramatically. For *spurious* MAR, i.e. the conditional probabilities remain the same over the level of the incomplete variable, the performance is comparable to that under MCAR. In fact, this would amount to one of the special cases under which complete case analysis would be more efficient than multiple imputation Van Buuren (2012), p. 48: the missingness does not depend on the incomplete variable. This property might be useful in practice, but considering it as a condition to evaluate performance under MAR missingness is useless.

Although one cannot definitively verify if the missingness is random - after all, for every MNAR mechanism there is a MAR mechanism with equal fit (Molenberghs et al. 2008) - it can be argued that MNAR is the more likely mechanism for real life missingness scenarios. In such cases, an indication of the validity of the obtained inference, given that the assumed missingness mechanism is suspected to be invalid, may be obtained by performing sensitivity analysis see e.g. Molenberghs et al. (2014), part 5.

Performance evaluation

[Add bridge from last section. And explain how this part is subdivided: first general characteristics of the imputations such as convergence, then performance measures.]

[Performance of different imputation methods may be influenced by preconditions of the methods themselves. All imputation models should be congenial, but this is difficult to test to often ignored even in single imputation endeavor, let alone in simulation studies with 100s of imputations. Same goes for convergence: FCS is iterative and requires algorithmic convergence, but this is typically evaluated through visual inspection which is unfeasible in simulation studies. Ignoring congeniality may yield sub-optimal imputations, which may disadvantage certain methods compared to others.]

Moreover, the performance of imputation procedures on distributional properties is often ignored in simulation studies, and even though the estimates on the analysis level may be justified, some methods can yield imputations that may seem completely invalid to applied researchers. For example, one could very accurately estimate average human height by filling in negative values and values that are unrealistically large. While the obtained inference could still be valid under such imputations, the plausibility of the imputed values given the observed data should be under scrutiny. Rather, one would prefer an imputation technique to yield both valid inference and plausible imputations. It should be studied if an imputation method is prone to deliver such impractical results, and if so, under what conditions.

[Then onto the performance measure part.]

The evaluation criteria used to assess imputation performance vary from one developer to another. This is not surprising as people from different fields could have a different focus on the problem at hand. But the choice of performance measure(s) may inadvertently distort statistical properties of the imputed data.

~~However, because multiple imputation was designed as a mode for statistical inference, a minimal set of statistical properties should at least be evaluated. We highlight four aspects of evaluation that should be considered when evaluation imputation routines.~~

Developers often only inquire about the 'accuracy' (i.e. how well can the method reproduce the original data). [Add RMSE at cell level here?] The goal of multiple imputation is not to reproduce the data, but to allow for obtaining valid inference given that the data are incomplete. We are interested in the correct answer to the research question; not in the truth itself. This means that, given the framework provided by Donald B. Rubin (1987), statistical properties such as bias, confidence intervals and the coverage rate of the confidence intervals should be studied. After all, the 95% confidence interval should contain the 'true' value at least 95 out of 100 times Neyman (1934), p. 591. [So, what could go wrong by focusing on accuracy exclusively?]

[Add choice of analysis model/estimand here, since performance measure are defined wrt the estimand, e.g., focusing on univariate estimates may favor invalid ad hoc methods.]

[Add problems with RMSE here.]

Sampling variation has always been an essential part of the evaluation of multiple imputation methodology. However, in order to obtain information about a method's ability to handle the missing data problem, or to objectively compare methods on their ability to correct for missingness, it is not necessary to take sampling variation into account Vink and van Buuren (2014). After all, we are interested only in the missing data mechanism, and are not considering the noise induced by the sampling mechanism for evaluation in such studies.

Last, when the evaluator is on the verge of drawing conclusions about the performance of the imputation routine, the performance should be carefully qualified. Comparing the performance of an imputation routine given a population (or true) parameter allows for quantitative evaluation. However, in order to pose qualitative statements about the performance on simulated conditions, comparative methodology is required. For example, when claiming that imputation performance is unacceptable when deviations from normality become rather stringent, such performance is highly dependent on the simulation conditions that are used.

For a well-balanced judgement about the severity of the performance drop, comparative simulations with e.g. nonparametric models should be executed. A method may perform badly, but if it still outperforms every other approach, it may yet be of great practical relevance.

Illustration

[Add simulation study + results here!]

Suggested course of action

Step 1. Obtain truth

The evaluator must to decide whether there is a need for sampling variance in the simulations scheme. If sampling variance cannot be omitted from the simulation scheme, two approaches are possible:

- Model-based simulation: draw samples from a known probability distribution, such as the multivariate normal distribution. The theoretical parameters under which the samples are obtained will serve as the comparative truth during the simulations.
- Design-based simulation: sample data from a sufficiently large set. The parameters of the sufficiently large set will serve as the comparative truth during the simulations. This approach is often used in situations where a probability distribution is not available, or where real-life data structures are of interest. Applications of design-based simulation are often found in official statistics. A benefit of design-based simulation is the ability to use real-life observed data structures.

For situations where sampling variance is not of interest, the sampling process can be omitted and a single complete dataset can simply be obtained by model- or design-based approaches. The parameters of that single complete set will serve as comparative truth. This process is computationally convenient, because only a single complete datasets has to be considered during all of the simulations. One needs to realize, however, that the conventional pooling rules (cf. Donald B. Rubin 1987, 76–77), do not apply for finite population inference and that alternative pooling rules need to be used (Raghunathan, Reiter, and Rubin 2003; Vink and van Buuren 2014).

Step 2. Induce missingness

First, one should always consider MCAR missingness, i.e. the scenario where the missing values are missing at random and the observed values are observed at random. Under MCAR, the statistical properties of the observed data given the missing data are known and any imputation routine that cannot at least mimic the performance of the observed data inference, should be deemed inefficient in the scope of the simulation.

Next, missing data should be induced conform a model that is dependent on the observed data. A straightforward technique for inducing different forms of univariate MAR missingness is described in Van Buuren (2012), p. 63; see Figure 2 and a generalization to multivariate MAR missingness can be found in Buuren et al. (2006), Appendix B, and Brand (1999), par. 5.2.3. If the missingness is to be induced in longitudinal data, auto-regressive MAR models e.g. cf. Shara et al. (2015), model 2 and model 3, can be useful.

Third, it is advisable to investigate varying shapes of MAR missingness to achieve a more realistic indication of the robustness of the imputation performance across the range of random missingness. Given the simulated data-distributions, one random missingness model may be far more disastrous to the observed information than another model. This may influence the performance of some (but not necessarily all) imputation routines For example, inference from hot-deck techniques such as predictive mean matching (Roderick JA Little 1988; Donald B. Rubin 1986) may be more severely impacted by large amounts of one-tailed missingness than inference from parametric techniques. It would be a shame to overlook such results due to the focus on a single MAR mechanism.

Fourth, the amount of missingness must be varied. Remember that missingness is only ignorable under MAR when the parameter of the data is distinct and a-priori independent from the parameter of the missing data
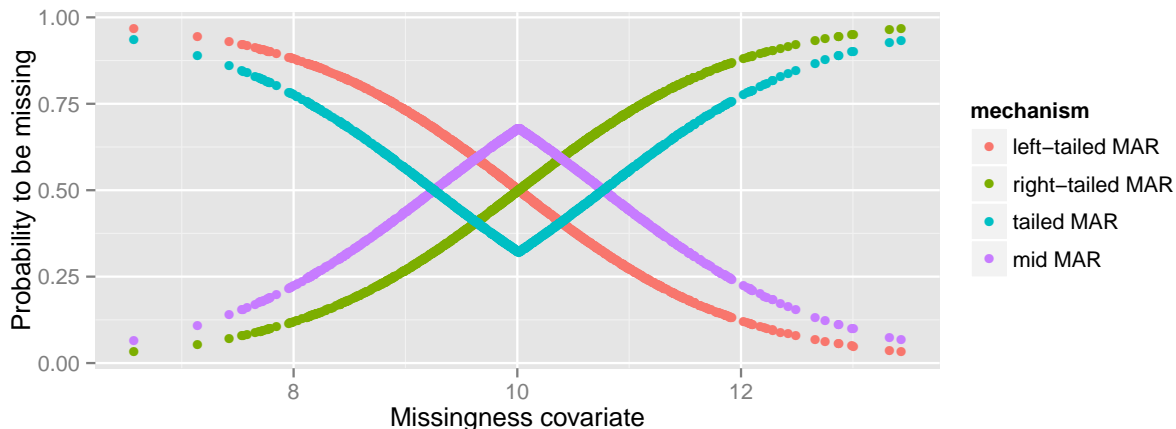
Figure 2: Examples of four random missingness mechanisms that yield approximately 50 percent missingness. For negatively correlated missingness covariates the resulting mechanisms are reversed.

process. Under MAR missingness we assume that we may use the observed data to make inferences about the joint (observed and unobserved) data.

The dependency of the procedure on the assumption under which we obtain inference is only influenced by the amount of missingness. If there is no missingness - or if there is no data, for that matter - the inference does not depend on the assumption. Alternatively, the validity of assumptions become increasingly important when the missingness increases. Since we control the MAR mechanism, the assumption under which we may solve the missing data problem should hold and it is only fair to assess performance under stringent missingness conditions. I therefor propose, for all mechanisms, to evaluate at least the following univariate missingness percentages when evaluating imputation routines:

- 10% missingness: Depending on the size of the data, this percentage can be considered as a lower bound of realistic evaluation. Anything less than 10% may be of little influence on the true data inference. Performance of a missing data method should at least be acceptable for most missing data problems.
- 25% missingness: This is a fair amount of missingness and will, depending on the observed data information, have a noticeable influence on the completed data inference. When compared to the condition with 10% missingness, the inference obtained under 25% missingness should be less certain (i.e. confidence/credibility interval width should increase), but estimates should still be properly covered and the statistical properties of the missing data method should be sound. In practice, at least to my experience in social sciences and official statistics, 25% univariate missingness can easily be considered as a realistic missingness percentage.
- 50% missingness: Performance under 50% percent simulated missingness will most likely be impacted severely. Depending on how the missingness mechanism interacts with the simulated data, some imputation techniques may yield estimates that are under-covered such that the completed data inference should not be deemed valid anymore. If a method yields acceptable inference under 50% MAR missingness, we can determine that the statistical properties of the imputation methodology are sound.

Although I limit the focus here to ignorable non-response, the above suggested proportions are equally applicable to simulations under non-ignorable non-response.

Fifth, the used missingness mechanism should be detailed, either graphically or written as a function of the data. Often, when inducing missingness, authors remain vague about the actual missingness mechanism under investigation and, even worse, some authors only report something like

> We generated missing data following a MAR missingness mechanism.

This should be considered unacceptable as claims about the validity of the multiple imputation inference depends heavily on the simulated missingness mechanism.

Step 3. Evaluate performance

It is wise to evaluate the performance of complete case analysis (aka list-wise deletion) in all simulated conditions. We know the theoretical properties of complete case analysis, which makes the technique useful as a point of departure when evaluating multiple imputation performance. The evaluation criteria may depend on the estimand. When descriptive statistics are the goal and when statistical inference would not be of interest, bias of the estimates would still apply, but standard errors are generally ignored. Chances are that for those who focus on descriptive statistical applications, multiple imputation would not be the mode of choice. In general, I would say that each multiple imputation routine should be evaluated on at least the following points:

- *Bias:* Results should preferable be unbiased. However, the way bias is considered can greatly influence the interpretation of the results. For example, a negligible absolute bias for a parameter for which the true value is zero, would yield infinite bias when relative bias is considered. The way bias is computed should therefore be carefully chosen and described.

- *Interval width:* The way the confidence interval is calculated should be described. Wider intervals are associated with more uncertainty and the more narrow interval that is still properly covered indicates a sharper inference. However, inference from a wider interval that is properly covered is to be considered more valid when compared to a more narrow interval that is not properly covered anymore.

- *Coverage:* Coverage of a 95% interval should in theory be ≥ 95, where a coverage of 95% would be most efficient. Under-coverage (when estimation is too liberal) may be an indication of invalid inference, while over-coverage (when estimation is too conservative) tells us that efficiency could still be gained.

- [Add RMSE 2x]

- *Distributional characteristics:* In practice, the distribution of the incomplete data may differ greatly from the observed data. Under anything but the MCAR assumption, this can be expected. When evaluating imputations, the distributional shapes should be checked and diagnostic evaluations should be performed see Abayomi, Gelman, and Levy (2008) for an detailed overview of diagnostic evaluation for multivariate imputations. When anomalies are found, and if the imputation method is valid, there should be an explanation, especially in the controlled environment of a properly executed simulation study.

- *Plausibility of the imputed values:* Plausible imputations - imputations that could be real values if they had been observed - are not a necessary condition for obtaining valid inference. However, in practice, especially when the imputer and the analyst are different persons, plausibility of imputations may be a desired property. When evaluating imputation routines, the evaluator should mention whether the routine is prone to deliver implausible values. [Add PPC]

- *Convergence of the algorithm:* Most contemporary imputation techniques rely on iterative algorithms, such as the Gibbs sampler, where some algorithms are critically considered to be possibly incompatible Gibbs samplers, PIGS, Li, Yu, and Rubin (2012). The convergence of all iterative algorithms should always be considered and if non-convergence is suspected, the inference resulting from the imputations should not be considered. [Extend this part!]

- [Add PPC (Mingyang) (Nguyen, Carlin, and Lee 2017), (Zhao 2022)]

For evaluations of model-based simulations, it could be convincing to demonstrate a method's applicability to real-world missing data problems. This can, for example, be achieved by obtaining and imputing a fully observed set of variables, wherein the missingness is mimicked from a similar, incomplete set. Alternatively, an incomplete set could be obtained and truth could be established by removing the incomplete cases from the data. However, the real-world missingness would then be omitted.

Discussion

This document is aimed at establishing a common ground for the evaluation of imputation routines. Such a common ground would be the basis of a standardized evaluation. This allows for fair and efficient comparisons between imputation techniques. Ultimately, it would be desirable to evaluate every imputation routine against the same standardized set in order to quantify the statistical properties across imputation routines. If properly executed, this would allow for careful matching of imputation methodologies to new missing data problems.

Appendix A: Simulation setup

Let $Y = (Y_{obs}, Y_{mis})$ be an incomplete semi-continuous variable, where Y_{obs} and Y_{mis} denote the observed values and the missing values in Y, respectively. Further, $X = (X_1, ..., X_k)$ is a set of k fully observed covariates, where X_{obs} and X_{mis} correspond to the observed an missing parts in Y and where X_j would indicate the jth variable in X, with $j = 1, \ldots, k$. Finally, let R be a response indicator that is 1 if Y is observed and 0 if Y is missing.

To prove the point, I generated a single data set from the multivariate normal distribution with means

$$\mu = \begin{matrix} Y \\ X_1 \\ X_2 \end{matrix} \begin{pmatrix} 5 \\ 10 \\ 10 \end{pmatrix},$$

and covariance matrix

$$\Sigma = \begin{matrix} Y \\ X_1 \\ X_2 \end{matrix} \begin{matrix} Y & X_1 & X_2 \\ \begin{pmatrix} 1 & 0.8 & 0 \\ 0.8 & 1 & 0 \\ 0 & 0 & 1 \end{pmatrix} \end{matrix}.$$

Three incomplete sets were then generated, wherein random missingness was imposed in Y in three ways. First, I left the observed data *observed at random* to impose MCAR missingness following the mechanism

$$P(R = 0|Y_{obs}, Y_{mis}, X_j) = P(R = 0) = .50.$$

Next, I created random missingness in Y according to the following MAR missingness mechanism

$$P(R = 0|Y_{obs}, Y_{mis}, X_j) = P(R = 0|Y_{obs}, X_j),$$

by using a random draw from a binomial distribution of the same length as Y and of size 1 with missingness probability equal to the inverse logit

$$P(R = 0) = \frac{e^a}{(1 + e^a)}.$$

where $a = (\mu_{X_j} - X_{ij})/\sqrt{\mathbb{E}[X_{ij} - \mu_{X_j}]}$ in order to create approximately 50% right-tailed MAR missingness. I followed this procedure twice; once for each X_j.

The three resulting incomplete data sets were then imputed with package `mice` Van Buuren and Groothuis-Oudshoorn (2011), v2.25, in `R` Team (2015), v3.2.3, with Bayesian linear regression imputation (`mice.impute.norm`) as the imputation routine. The completed-data analyses were combined into a single inference following the rules described in Vink and van Buuren (2014).

```
## R version 4.1.2 (2021-11-01)
## Platform: x86_64-w64-mingw32/x64 (64-bit)
## Running under: Windows 10 x64 (build 19042)
##
## Matrix products: default
```

```
##
## locale:
## [1] LC_COLLATE=Dutch_Netherlands.1252  LC_CTYPE=Dutch_Netherlands.1252
## [3] LC_MONETARY=Dutch_Netherlands.1252 LC_NUMERIC=C
## [5] LC_TIME=Dutch_Netherlands.1252
##
## attached base packages:
## [1] stats     graphics  grDevices utils     datasets  methods   base
##
## other attached packages:
## [1] dplyr_1.0.8
##
## loaded via a namespace (and not attached):
##  [1] rstudioapi_0.13 knitr_1.37       magrittr_2.0.2  tidyselect_1.1.2
##  [5] R6_2.5.1         rlang_1.0.1     fastmap_1.1.0   fansi_1.0.2
##  [9] highr_0.9        stringr_1.4.0   tools_4.1.2     xfun_0.29
## [13] utf8_1.2.2       DBI_1.1.2       cli_3.2.0       htmltools_0.5.2
## [17] ellipsis_0.3.2   assertthat_0.2.1 yaml_2.3.5     digest_0.6.29
## [21] tibble_3.1.6     lifecycle_1.0.1 crayon_1.5.0    purrr_0.3.4
## [25] vctrs_0.3.8      glue_1.6.2       evaluate_0.15  rmarkdown_2.11
## [29] stringi_1.7.6    compiler_4.1.2  pillar_1.7.0    generics_0.1.2
## [33] pkgconfig_2.0.3
```

References

Abayomi, Kobi, Andrew Gelman, and Marc Levy. 2008. "Diagnostics for Multivariate Imputations." *Journal of the Royal Statistical Society: Series C (Applied Statistics)* 57 (3): 273–91.

Brand, Jaap. 1999. "Development, Implementation and Evaluation of Multiple Imputation Strategies for the Statistical Analysis of Incomplete Data Sets." PhD thesis, Erasmus University Rotterdam.

Buuren, S. Van, J. P. L. Brand, C. G. M. Groothuis-Oudshoorn, and D. B. Rubin. 2006. "Fully Conditional Specification in Multivariate Imputation." *Journal of Statistical Computation and Simulation* 76 (12): 1049–64. https://doi.org/10.1080/10629360600810434.

Li, Fan, Yaming Yu, and Donald B Rubin. 2012. "Imputing Missing Data by Fully Conditional Models: Some Cautionary Examples and Guidelines." *Duke University Department of Statistical Science Discussion Paper* 1124.

Little, Roderick JA. 1988. "Missing-Data Adjustments in Large Surveys." *Journal of Business & Economic Statistics* 6 (3): 287–96.

Little, Roderick J, Ralph D'Agostino, Michael L Cohen, Kay Dickersin, Scott S Emerson, John T Farrar, Constantine Frangakis, et al. 2012. "The Prevention and Treatment of Missing Data in Clinical Trials." *New England Journal of Medicine* 367 (14): 1355–60.

Little, and Rubin. 2020. *Statistical Analysis with Missing Data, Third Edition | Wiley Series in Probability and Statistics*.

Liu, Dawei, Hanne I. Oberman, Johanna Muñoz, Jeroen Hoogland, and Thomas P. A. Debray. 2021. "Quality Control, Data Cleaning, Imputation." In *Clinical Applications of Artificial Intelligence in Real-World Data*. https://arxiv.org/abs/2110.15877.

Meng, Xiao-Li. 1994. "Multiple-Imputation Inferences with Uncongenial Sources of Input." *Statistical Science* 9 (4): 538–58. https://doi.org/10.1214/ss/1177010269.

Molenberghs, Geert, Caroline Beunckens, Cristina Sotto, and Michael G Kenward. 2008. "Every Missingness Not at Random Model Has a Missingness at Random Counterpart with Equal Fit." *Journal of the Royal Statistical Society: Series B (Statistical Methodology)* 70 (2): 371–88.

Molenberghs, Geert, Garrett Fitzmaurice, Michael G Kenward, Anastasios Tsiatis, and Geert Verbeke. 2014. *Handbook of Missing Data Methodology*. CRC Press.

Morris, Tim P., Ian R. White, and Michael J. Crowther. 2019. "Using Simulation Studies to Evaluate Statistical Methods." *Statistics in Medicine* 38 (11): 2074–2102. https://doi.org/10.1002/sim.8086.

Neyman, Jerzy. 1934. "On the Two Different Aspects of the Representative Method: The Method of Stratified Sampling and the Method of Purposive Selection." *Journal of the Royal Statistical Society* 97 (4): 558–625. https://doi.org/10.2307/2342192.

Nguyen, Cattram D., John B. Carlin, and Katherine J. Lee. 2017. "Model Checking in Multiple Imputation: An Overview and Case Study." *Emerging Themes in Epidemiology* 14 (1): 8. https://doi.org/10.1186/s12982-017-0062-6.

Raghunathan, Trivellore E, Jerome P Reiter, and Donald B Rubin. 2003. "Multiple Imputation for Statistical Disclosure Limitation." *Journal of Official Statistics* 19 (1): 1.

Rubin, Donald B. 1986. "Statistical Matching Using File Concatenation with Adjusted Weights and Multiple Imputations." *Journal of Business & Economic Statistics* 4 (1): 87–94.

Rubin, Donald B. 1976. "Inference and Missing Data." *Biometrika* 63 (3): 581–92. https://doi.org/10.2307/2335739.

———. 1987. *Multiple Imputation for Nonresponse in Surveys.* Wiley Series in Probability and Mathematical Statistics Applied Probability and Statistics. New York, NY: Wiley.

Shara, Nawar, Sayf A Yassin, Eduardas Valaitis, Hong Wang, Barbara V Howard, Wenyu Wang, Elisa T Lee, and Jason G Umans. 2015. "Randomly and Non-Randomly Missing Renal Function Data in the Strong Heart Study: A Comparison of Imputation Methods." *PloS One* 10 (9): e0138923.

Team, R Core. 2015. *R: A Language and Environment for Statistical Computing.* Manual. Vienna, Austria: R Foundation for Statistical Computing.

Van Buuren, Stef. 2012. *Flexible Imputation of Missing Data.* CRC press.

Van Buuren, Stef, and Karin Groothuis-Oudshoorn. 2011. "Mice: Multivariate Imputation by Chained Equations in R." *Journal of Statistical Software* 45 (1): 1–67. https://doi.org/10.18637/jss.v045.i03.

Vink, Gerko, and Stef van Buuren. 2014. "Pooling Multiple Imputations When the Sample Happens to Be the Population," September.

Zhao, Yang. 2022. "Diagnostic Checking of Multiple Imputation Models." *AStA Advances in Statistical Analysis*, January. https://doi.org/10.1007/s10182-021-00429-1.